big data and cognitive computing

MDPI

Article

Listening to Patients: Advanced Arabic Aspect-Based Sentiment Analysis Using Transformer Models Towards Better Healthcare

Seba AlNasser * and Sarab AlMuhaideb

Department of Computer Science, College of Computer and Information Sciences, King Saud University, P.O. Box 266, Riyadh 11362, Saudi Arabia; salmuhaideb@ksu.edu.sa
* Correspondence: 444203397@student.ksu.edu.sa

Abstract: Patient satisfaction is a key measure of the quality of healthcare, directly impacting the success and competitiveness of healthcare providers in an increasingly demanding market. Traditional feedback collection methods often fall short of capturing the full spectrum of patient experiences, leading to skewed satisfaction reports due to patients' reluctance to criticize services and the inherent limitations of survey designs. To address these issues, advanced Natural Language Processing (NLP) techniques such as aspect-based sentiment analysis are emerging as essential tools. Aspect-based sentiment analysis breaks down the feedback text into specific aspects and evaluates the sentiment for each aspect, offering a more nuanced and actionable understanding of patient opinions. Despite its potential, aspect-based sentiment analysis is under-explored in the healthcare sector, particularly in the Arabic literature. This study addresses this gap by performing an Arabic aspect-based sentiment analysis on patient experience data, introducing the newly constructed Hospital Experiences Arabic Reviews (HEAR) dataset, and conducting a comparative study using Bidirectional Embedding Representations from Transformers (BERT) combined with machine learning classifiers, as well as fine-tuning BERT models, including MARBERT, ArabicBERT, AraBERT, QARiB, and CAMeLBERT. Additionally, the performance of GPT-4 via OpenAI's ChatGPT is evaluated in this context, making a significant contribution to the comparative study of BERT with traditional classifiers and the assessment of GPT-4 for aspect-based sentiment analysis in healthcare, ultimately offering valuable insights for enhancing patient experiences through the use of AI-driven approaches. The results show that the joint model leveraging MARBERT and SVM achieves the highest accuracy of 92.14%, surpassing other models, including GPT-4, in both aspect category detection and polarity tasks.

Keywords: natural language processing; aspect-based sentiment analysis; patient satisfaction

Citation: AlNasser, S.; AlMuhaideb, S. Listening to Patients: Advanced Arabic Aspect-Based Sentiment Analysis Using Transformer Models Towards Better Healthcare. *Big Data Cogn. Comput.* **2024**, *8*, 156. https://doi.org/10.3390/bdcc8110156

Academic Editor: Alberto Fernandez Hilario

Received: 26 September 2024
Revised: 6 November 2024
Accepted: 12 November 2024
Published: 14 November 2024

1. Introduction

Healthcare providers rely on patient feedback to improve the patient experience and enhance satisfaction, as it is one of the core indicators used to measure the quality of service and success of healthcare providers. One of the traditional ways to collect patient feedback is through surveys; however, the results are sometimes skewed towards indicating high satisfaction levels. This issue results from patients' unwillingness to criticize healthcare services and is partially due to the design of the surveys. Therefore, using methods that allow patients to describe their experiences of healthcare services from their own perspective typically gives a more realistic view [1]. Social media platforms (e.g., Twitter and Google Maps) could be valuable information sources, as patients tend to freely express their feelings and opinions about healthcare services, including their feedback on the healthcare provider, doctors, treatment, prices, and appointments.

AlMuhaideb et al. [2] collected a dataset comprising patient experiences in Arabic from the Twitter platform and conducted sentiment analysis using a deep learning model. However, sentiments were analyzed with respect to the healthcare service as a whole, which cannot accurately reflect the specific aspects mentioned by the patient, mainly as a tweet

can express different opinions towards different aspects. Aspect-based sentiment analysis is an effective tool for identifying the aspects discussed in a given text and determining the opinions or sentiments of these identified aspects. Therefore, the results obtained from aspect-based sentiment analysis are helpful for both patients in choosing healthcare providers and healthcare providers in enhancing the patient experience. Despite the crucial need for such advanced analyses, only a few studies have explored the use of aspect-based sentiment analysis in the healthcare domain, especially with respect to the Arabic literature. Therefore, in the study reported herein, the authors seek to classify sentiments on a more fine-grained level through the use of aspect-based sentiment analysis, which aims to extract the aspects and their related polarities from the patient experience, thus providing a comprehensive view of the strengths and weaknesses of the provided healthcare service and enabling the healthcare provider to make targeted improvements.

The task of aspect-based sentiment analysis can be divided into four sub-tasks [3]: aspect term extraction, aspect term polarity, aspect category detection, and aspect category polarity. Aspect term extraction is the task of identifying all aspect terms present in a given text. Aspect term polarity is the task of determining the polarity of each aspect term presented in a given text, assuming that the present aspect terms are provided. Aspect category detection is the task of identifying the aspect categories discussed in a given text from a pre-defined list of aspect categories. Aspect category polarity is the task of determining the polarity of each discussed category in a given text. This work focuses on two aspect-based sentiment analysis sub-tasks—namely, aspect category detection and aspect category polarity—which involve identifying the aspect categories discussed in a given text from a pre-defined list of aspect categories and determining the polarity of each. The word detection in this context implies recognizing the presence of a label. As the text may include more than one aspect category, this problem can be considered a multi-label classification problem.

The authors aimed to answer the following research questions: (1) What are patients' attitudes towards different aspects of the healthcare services provided in Saudi Arabia? (2) How accurate are shallow machine learning classifiers using BERT as word embeddings in the aspect category detection and polarity sub-tasks? (3) How well do generative models (i.e., GPT-4) perform in aspect-based sentiment analysis, compared with BERT-based models?

In summary, the anticipated contributions of the manuscript are as follows:

- Introduction of a Novel Arabic Patient Experience Dataset: The study introduces a Google Maps-based dataset of Arabic patient reviews on healthcare providers in Saudi Arabia (i.e., Hospital Experiences Arabic Reviews; HEAR) and annotates the HoPE-SA dataset for aspect detection and polarity tasks.
- Comparative Study of BERT-based Models with Machine Learning Classifiers: The study evaluates fine-tuned BERT models with traditional classifiers (Support Vector Machine (SVM) [4], Random Forest (RF) [5], and neural networks [6]) for aspect-based sentiment analysis.
- Performance Analysis of Joint and Two-Stage Models: The study compares joint and two-stage models, as well as GPT-4, revealing that the joint model significantly outperforms the two-stage approach and slightly surpasses GPT-4 in aspect detection and sentiment polarity tasks.

The rest of this paper is organized as follows: Section 2 presents the related work in the field of aspect-based sentiment analysis. Section 3 provides a detailed description of the process of collecting, annotating, and analyzing Arabic tweets and reviews with respect to several aspects. The results and findings are presented and discussed in Sections 5 and 6, and the paper is concluded in Section 7.

2. Related Work

2.1. Classical Machine Learning Approaches

Alassaf et al. [7] adopted classical machine learning methods for aspect-based sentiment analysis, for which they collected Arabic tweets related to the education sector in Saudi Arabia to perform aspect-based sentiment analysis. They used the Term Frequency–Inverse Document Frequency (TF-IDF) feature extraction method, a hybrid feature selection method consisting of one-way ANOVA based on F-values and regularization, and a Support Vector Machine (SVM) [4] for the classification task. The use of a hybrid feature selection method enhanced the model's performance and its ability to overcome the high-dimensionality problem in text classification. However, the dataset used in this study was limited to one university in Saudi Arabia. Almasaud et al. [8] created an Arabic multi-aspect and multi-sentiment dataset of Google Maps restaurant reviews. They used four machine learning models: Naïve Bayes (NB), Support Vector Classification (SVC), linear SVC, and Stochastic Gradient Descent (SGD) [9]. The best results were achieved by the SVC and linear SVC models. The machine learning-based approaches in the mentioned studies relied on hand-crafted features, which are surpassed by the state-of-the-art automatic feature extraction empowered by deep learning methods.

2.2. Convolutional and Recurrent Neural Network Approaches

For aspect term-related tasks, Al-Smadi et al. [10] proposed two models based on Long Short-Term Memory (LSTM) [11] to solve Arabic aspect term extraction and polarity separately. The aspect extraction model consists of three layers: an embedding layer using Word2vec [12] and FastText [13], a Bidirectional Long Short-Term Memory (BiLSTM) layer, and a Conditional Random Fields (CRFs) [14] layer. Meanwhile, the sentiment polarity model consists of LSTM, attention, and Softmax layers. They evaluated the proposed model on the Arabic Hotel Reviews dataset part of the SemEval 2016 Task 5 dataset. The results showed that using FastText character embeddings yielded better results, compared to Word2vec word embeddings. In contrast, Kuppusamy et al. [15] proposed a hybrid deep learning model that jointly tackles the aspect term extraction and polarity tasks by concatenating Convolutional Neural Networks (CNNs) [16] and BiLSTM models. Han et al. [17] utilized Bidirectional Gated Recurrent Units (BiGRUs) to perform aspect term polarity assessment for drug reviews. They proposed a pre-training and multi-task learning model based on double BiGRU and a dataset called SentiDrugs. The proposed model is divided into an embedding layer, double BiGRU layer, attention layer, and Softmax layer. They constructed the SentiDrugs dataset by randomly selecting 4200 reviews shorter than 200 words on the effectiveness and side effects from Druglib.com. The results showed that the proposed model outperformed the baseline methods, including LSTM and BiGRU.

For aspect category-related tasks, Sivakumar et al. [18] proposed the use of an LSTM with a fuzzy logic model to classify consumer review sentences under various aspects with four different labels, namely, highly negative, negative, positive, and highly positive. Initially, the dataset was grouped according to geographical location. The proposed model was trained and tested separately for every dataset, based on country. Records from only the most recent three years were considered for training and testing the developed system. Thus, the current trends can also be considered for making decisions on input customer reviews. Word embedding using Continuous Bag of Words (CBOW) [12] was used to extract aspects from sentences. The sentiment score was generated using an LSTM, following which the score was passed to the fuzzy logic system for classification into one of four categories. Gao et al. [19] constructed a hybrid model consisting of a CNN and a BiGRU for Chinese aspect category detection and polarity. The use of BiGRU, along with a CNN, alleviated the vanishing gradient problem and enhanced the model's ability to learn sequence information.

Al-Dabet et al. [20] proposed two deep learning models to address the Arabic aspect category detection and polarity tasks separately. The aspect category detection model is decomposed into independent binary classifiers, each trained on a single aspect category.

It consists of a CNN and a stacked independent LSTM [21]. The aspect category polarity model consists of five primary modules: an input module (skip-gram for character level and CBOW for word level), a stacked bidirectional independent LSTM module, a position-weighting mechanism module, a multiple attention mechanism module, and an output module. They conducted experiments on the Arabic Hotel Reviews dataset, part of the SemEval 2016 Task 5 dataset, and the results demonstrated that the proposed models outperformed the baseline and other models, with the first model achieving an F1 measure of 58.08%, and the second model achieving an accuracy measure of 87.31%.

2.3. Transformer-Based Approaches

For aspect term-related tasks, Apostol et al. [22] proposed a heterogeneous ensemble model to solve English aspect-based sentiment analysis in terms of two sub-tasks: aspect term extraction and aspect term polarity. The proposed model consists of a linear model, BiLSTM, CNN-BiLSTM, and pre-trained and fine-tuned Bidirectional Embedding Representations from Transformers (BERT) [23] and Bidirectional and Auto-Regressive Transformers (BART) [24] as word embeddings. They evaluated the proposed model on the SemEval 2016 Task 5 Restaurants dataset and the Multi-Aspect Multi Sentiment dataset. The accuracies for the task of aspect term extraction were 99.96% and 99.95% for the two datasets, respectively. For the aspect term polarity task, the accuracies were 99.74% and 99.87% for the two datasets, respectively. To solve the problem of aspect detection and sentiment classification jointly, Rani et al. [25] proposed a multi-task learning-based dual BiLSTM model for aspect-based sentiment analysis of drug reviews. The first layer of the proposed model includes contextual embeddings initialized using BERT, and aspect-specific representations are generated by multi-head self-attention to emphasize different parts of the input sequence. The second layer of the model consists of a dual BiLSTM—one for the BERT embeddings and the other for aspect-specific representations. The final layer of the model is the Softmax layer for the classification task.

Similarly, Chouikhi et al. [26] utilized transfer learning for Arabic aspect-based sentiment analysis. They used an Arabic-based BERT stacked with a CRF layer. They used an annotation scheme that considers the aspect term and its polarity, and jointly solved the aspect term extraction and aspect term polarity. This model benefits from the characteristics of CRF, which can provide certain constraints on the output label that adhere to the labeling scheme. Furthermore, Li et al. [27] formulated the problem of aspect term extraction and aspect term polarity as a sequence labeling problem and solved both tasks jointly. They used BERT as word embeddings along with four different models in the classification layer: a fully connected layer, Gated Recurrent Units (GRUs) [28], a self-attention network, and a CRF. The results on the SemEval 2014 Task 4 dataset showed that the best results were achieved by GRU, with an F1-score of 61.12%.

Abdelgwad et al. [29] explored the Arabic aspect term polarity as a sentence-pair problem. They proposed an Arabic BERT-based model with a linear classification layer. The model receives two sentences—the review sentence and the aspect terms—and the task is to determine the sentiments towards each aspect. Their results demonstrated that the proposed model achieved better results than many previous Arabic deep learning models through simply adding a linear layer on top of BERT. Fadel et al. [30] proposed an Arabic aspect term extraction model using a stacked embeddings layer consisting of AraBERT [31] and Flair [32] embeddings, stacked Recurrent Neural Network (RNN)-based models [33], and CRF layers. They employed two versions of the model using BiLSTM and BiGRU. The results indicated that the BiLSTM version of the model outperformed BiGRU.

For aspect category-related tasks, Chang et al. [34] jointly tackled aspect category extraction and polarity, where they investigated customer satisfaction through aspect-level sentiment analysis and visual analytics. They collected flight reviews from Tripadvisor.com in order to measure the impact of COVID-19 on passenger travel sentiments in several aspects. They selected the top 12 airlines from Tripadvisor.com and collected their reviews from January 2016 to August 2020. They split the reviews into two groups: before and

during the pandemic. They then classified the reviews based on the star ratings of eight aspect categories into negative (1–3) or positive (4–5). They used BERT followed by multi-head attention and a fully connected layer. The model achieved a 60% F1-score across the eight aspect categories. Meanwhile, Hoang et al. [35] proposed three models for aspect category detection and aspect category polarity. The first model is the aspect category classifier, which predicts whether the aspect is related to the text. The second model is the sentiment classifier, which predicts the sentiment polarity of the text and aspect. The third is a combined model that takes the text and aspect as inputs and predicts the polarity if the aspect is related. All three proposed models used BERT and formulated the problem as sentence-pair classification. The results showed that the aspect classifier outperformed the combined classifier, whereas the combined classifier outperformed the sentiment analysis.

One of the challenges faced when conducting aspect-based sentiment analysis is the limited availability of labeled data points as highlighted by Shim et al. [36], who proposed a label-efficient training scheme to overcome this challenge. For the construction of an auxiliary sentence, Sun et al. [37] investigated four methods to construct an auxiliary sentence and transform aspect-based sentiment analysis into a sentence-pair classification task. They fine-tuned the pre-trained BERT model on an aspect-based sentiment analysis task with a classification layer and Softmax function. They evaluated the model on the SemEval 2014 Task 4 dataset, and the results showed that the BERT pair outperformed the single BERT model. Li et al. [38] utilized two methods to construct an auxiliary sentence and used a gating mechanism with context-aware aspect embeddings to enhance and control the BERT representation for aspect-based sentiment analysis. The main idea of context-aware embeddings is to select highly correlated words from the context. The gating layer controls the propagation of sentiment features from the BERT representation with context-aware embeddings. The results showed that using context-aware embeddings enhances the model performance. Additionally, the use of the BERT pair outperformed the single BERT.

For aspect term and category-related tasks, Al-jarrah et al. [39] performed Arabic aspect term extraction, aspect category detection, and aspect term polarity. They collected and annotated Arabic tweets about food delivery service reviews with different Arabic dialects for aspect-based sentiment analysis. They used two deep learning models (BiLSTM-CRF and LSTM), two transformer-based models (GigaBERT [40] and AraBERT), and a classical machine learning model (SVM with TF-IDF) for feature extraction. The results demonstrated that the transformer-based models outperformed the other models, as AraBERT and GigaBERT have been pre-trained on a large amount of data, thus enhancing the learned language representation. Bensoltane et al. [41] performed Arabic aspect term extraction and aspect category detection. For the aspect term extraction, they used four models: AraBERT with linear layer, AraBERT with CRF layer, AraBERT with BiLSTM and CRF, and AraBERT with BiGRU and CRF. They used two approaches based on the BERT model for aspect category detection, where the first approach uses a fine-tuned AraBERT model, while the second uses AraBERT as a word embedding with CNN model. For the fine-tuning approach, they explored single-sentence classification and sentence-pair classification. Their results showed that AraBERT with BiGRU and CRF outperformed the other three models in the aspect term extraction task. For aspect category detection, AraBERT fine-tuned with sentence-pair classification outperformed other models. Furthermore, their results revealed that fine-tuning is suitable in the case of limited available data. Moreover, this work focused on extracting and detecting terms and categories but did not analyze the polarity of these terms and categories.

2.4. Critical Analysis

Based on the above-mentioned studies—except for that of Shim et al. [36], who performed English aspect category detection and polarity for healthcare-related program reviews—no study has explored the use of aspect-based sentiment analysis in the healthcare domain, especially in the context of the Arabic literature.

Based on those studies that explored the aspect term-related tasks, the studies that only involved the aspect term polarity task demonstrated that simply adding a linear layer on top of BERT enhanced the performance of the model. However, the studies that involved the aspect term extraction task showed that using models with more complex classification layers, such as LSTM, GRU, or CRF, yielded better results.

Based on the studies that explored the aspect category-related tasks, several works [35–38] have formulated the problem as a sentence-pair classification problem. Based on the studies of Sun et al. [37] and Li et al. [38], the present authors concluded that formulating the problem of aspect category detection and polarity as sentence-pair BERT outperforms single BERT when applied with the same settings. The superiority of sentence-pair BERT derives from the advantages of BERT in the sentence-pair classification task, given that it has been pre-trained on a masked language model and next-sentence prediction tasks. Additionally, constructing a dataset for sentence-pair classification expands the original dataset exponentially. As suggested by Al-jarrah et al. [39] and Li et al. [27], the transformer-based word embedding method, BERT, outperforms other frequency- and prediction-based word embedding methods, as it has been pre-trained on a large amount of data, enhancing the learned language representation, in addition to its ability to dynamically learn better contextual embeddings based on a given context. It is also worth mentioning that, except for ArabicBERT, GigaBERT, and AraBERT, no other BERT versions were utilized.

A notable deficiency in the existing literature can be observed with respect to aspect-based sentiment analysis, particularly in the context of the healthcare industry in the Arabic literature. Our review of 21 articles revealed that, aside from Shim et al. [36], who addressed English aspect category detection and polarity for healthcare-related program reviews, no studies have specifically examined this topic in the Arabic literature. This study seeks to address this gap by offering insights and analyses that can enhance understanding and improve methodologies in this important area of research. Table 1 presents a comparison of the retrieved literature in the context of aspect-based sentiment analysis.

Table 1. Comparative literature review of aspect-based sentiment analysis.

Study	Year	Approach	Dataset	Models	Type of Task	Language
Alassaf et al. [7]	2020	Classical ML	Arabic Tweets	TF-IDF, SVM	ABSA	Arabic
Almasaud et al. [8]	2023	Classical ML	Google Maps restaurant reviews	NB, SVC, Linear SVC, SGD	ABSA	Arabic
Al-Smadi et al. [10]	2019	RNN	Arabic Hotel Reviews	LSTM, FastText, BiLSTM, CRF	Aspect Term, Polarity	Arabic
Kuppusamy et al. [15]	2023	Hybrid DL	Not specified	CNN, BiLSTM	Aspect Term, Polarity	Not specified
Han et al. [17]	2020	RNN	SentiDrugs	BiGRU, Double BiGRU, Attention	Aspect Term, Polarity	English
Sivakumar et al. [18]	2021	RNN + Fuzzy Logic	Not specified	LSTM, CBOW, Fuzzy	Aspect Term	Not specified
Gao et al. [19]	2022	Hybrid DL	Chinese reviews	CNN, BiGRU	Aspect Category, Polarity	Chinese
Al-Dabet et al. [20]	2021	DL	Arabic Hotel Reviews	CNN, LSTM, Attention	Aspect Category, Polarity	Arabic
Apostol et al. [22]	2023	Transformer Ensemble	Arabic Hotel Reviews	BERT, BART, BiLSTM, CNN	Aspect Term, Polarity	English
Rani et al. [25]	2023	Multi-task Learning	Drug reviews	BERT, Dual BiLSTM, Attention	ABSA	English
Chouikhi et al. [26]	2023	Transformer + CRF	Arabic reviews	Arabic BERT, CRF	Aspect Term, Polarity	Arabic
Li et al. [27]	2019	Transformer + Sequence Labeling	SemEval 2014	BERT, GRU, Self-Attention, CRF	Aspect Term, Polarity	English
Abdelgwad et al. [29]	2022	BERT-based Model	Not specified	Arabic BERT, LR	Aspect Term, Polarity	Arabic
Fadel et al. [30]	2022	DL	Not specified	AraBERT, Flair, BiLSTM, BiGRU, CRF	Aspect Term	Arabic
Chang et al. [34]	2022	Transformer	Tripadvisor flight reviews	BERT, Multi-head Attention	Aspect Category	English
Hoang et al. [35]	2019	Transformer	Not specified	BERT-based Models	Aspect Detection, Polarity	Not specified
Shim et al. [36]	2021	Label-efficient Training	Not specified	Label-efficient Training	ABSA	Not specified
Sun et al. [37]	2019	Transformer + Sentence-Pair Classification	SemEval 2014	BERT	ABSA	English
Li et al. [38]	2020	Transformer + Context-Aware Embeddings	Not specified	BERT, Gating Mechanism	ABSA	Not specified
Al-Jarrah et al. [39]	2023	DL + Transformer	Arabic tweets	BiLSTM-CRF, LSTM, GigaBERT, AraBERT, SVM	Aspect Term, Category, Polarity	Arabic
Bensoltane et al. [41]	2022	DL + Transformer	Not specified	AraBERT, BiGRU, CRF, CNN	Aspect Term, Category	Arabic

3. Materials and Methods

3.1. Datasets

This work used the HoPE-SA [2] dataset, which contains 12,400 patient experience-related Arabic tweets from 14 healthcare organizations in Saudi Arabia. Each tweet has been labeled, according to its sentiment, as positive or negative. In this work, in order to perform aspect-based sentiment analysis, we annotated the dataset for aspect category detection and aspect category polarity tasks.

3.1.1. Aspect Category Identification

The first step in annotating the dataset for aspect category detection and polarity tasks is to identify the healthcare aspects discussed in the dataset. In order to do so, we first used the list of the most frequent words in the HoPE-SA dataset. Table 2 presents the top 20 most frequent words in the HoPE-SA dataset, with the potential categories in bold (words 6, 7, and 15).

Table 2. Top 20 frequent words in the HoPE-SA dataset [2], where the words in bold represent potential aspect categories.

No.	Word	Frequency	No.	Word	Frequency
1	Hospital	3242	11	**Doctor**	524
2	The Specialist	1478	12	AlHabib	499
3	The Hospital	1190	13	Complete	459
4	The Patients	1132	14	Faisal	457
5	The Patient	1006	**15**	**Service**	**439**
6	**The Doctor**	**895**	16	Hour	434
7	**Appointment**	**872**	17	Respond	431
8	Thanks	802	18	You Have	429
9	**The Appointments**	640	19	Unfortunately	417
10	Department	614	20	The Hour	409

However, the list of the most frequent words was not sufficient to identify the aspect categories, as one category may be discussed using many different terms (e.g., words 15, 16, 17, and 20). Thus, this work utilized GPT-4, through ChatGPT by OpenAI [42] in order to find the most frequent aspects and identify more fine-grained categories. We used the ChatGPT API to send and receive prompts and responses. Inspired by [43], the process started by prompt engineering to enable GPT-4 to assign aspects for each tweet. Figure 1 shows the designed prompt for this task in few-shot settings [43]. The input to the first example in the figure can be translated as follows: *"Sulaiman Al Habib Specialized Hospital, but unfortunately the appointments are very far away. I want an appointment and I'm trying, but there aren't any available"*. The second example tweet is translated as follows: *"Doctor Mohammed Shaker at Al Hammadi Hospital- Alsuwaidi (branch), has a lot of experience and is a very skilled artist with a light touch"*. The resulting aspects were sorted according to their frequency and, based on the results, five healthcare categories were initially identified.

```
As a social scientist, Your task is to extract discussed aspects of a
user tweet extracted from Twitter. Please assign aspects for each tweet.
Each aspect must not be longer than two words.

follow the following example:
######
example 1:
####
input:
مستشفى سليمان الحبيب التخصصي ، لكن للاسف مواعيده بعيده كثيير ، انا ابي
موعد وجالسه احاول لكن مافيه
output:
Appointments
####

example 2:
####
input:
دكتور محمد شاكر مستشفى الحمادي السويدي  عن تجربة ولله فنان ويده جذا خفيفه
output:
Doctors

####
######
```

Figure 1. The prompt for tweet aspect identification.

1. Medical staff;
2. Appointments;
3. Customer service;
4. Emergency services;
5. Pricing.

3.1.2. Pre-Labeling of HoPE-SA Dataset

After identifying the initial categories, this work also used GPT-4 for the aspect category polarity pre-labeling task. Initially, four classes were identified: positive, negative, neutral, and not mentioned. The latter refers to the aspect categories not discussed in the tweet. Figure 2 shows the designed prompt for this task in few-shot settings.

```
As a social scientist, Your task is to analyze the sentiment of a series
of user tweets extracted from Twitter. Please assign a sentiment score
from 0 to 3 for each category in the tweet, where 0 signifies negative
sentiment, 1 indicates positive sentiment,  2 indicates neutral
sentiment and 3 corresponds to not mentioned category. In situations
where the sentiment is difficult to definitively classify, please
provide your best estimation of the sentiment score.
The output must be in json format with only two keys the category and
the corresponding sentiment in the same line with the same order of
categories as the below example.

follow the following example:
######
example 1:
####
input:
مستشفى سليمان الحبيب التخصصي ، لكن للاسف مواعيده بعيده كثيير ، انا ابي
موعد وجالسه احاول لكن مافيه
output:
{"Pricing": 3, "Appointments": 0, "Medical Staff": 3, "Customer
Service": 3, "Emergency Services": 3}
####

example 2:
####
input:
دكتور محمد شاكر مستشفى الحمادي السويدي  عن تجربة والله فنان ويده جذا خفيفه
output:
{"Pricing": 3, "Appointments": 3, "Medical Staff": 1, "Customer
Service": 3, "Emergency Services": 3}

####
######
```

Figure 2. The prompt for aspect category polarity.

The pre-labeling results are provided in Table 3. These labels require human verification before being used for the training of aspect-based sentiment analysis models.

Additionally, the results showed that 173 instances were not classified into any of the initial categories. These instances were reviewed manually in the human verification stage, and mapped to one of the below cases:

- Case 1: The instance was misclassified by GPT-4.
- Case 2: The instance has no aspect discussed.
- Case 3: The aspect discussed in the instance does not belong to any of the initial categories.

The fact that only 173 instances out of the total 12,400 instances remained unclassified suggests that the initially identified categories effectively encompassed the aspects discussed in the dataset.

Based on the pre-labeling results, some categories had few instances, which may affect the performance of the aspect-based sentiment analysis models. Thus, we collected more data as an extension to the HoPE-SA dataset.

Table 3. Pre-labeling results for the HoPE-SA dataset using GPT-4 for aspect category polarity.

Polarity/Category	Negative	Positive	Neutral	Not Mentioned	Total
Medical Staff	3012	3091	204	6093	6307
Appointments	3047	165	73	9115	3285
Customer Service	6489	2863	202	2846	9554
Emergency Services	1603	274	186	10,337	2063
Pricing	748	96	30	11,526	874
Total	14,899	6489	695	39,917	22,083

3.1.3. Hospital Experiences Arabic Reviews (HEAR)

As an extension to the HoPE-SA dataset, we constructed an Arabic aspect-based sentiment analysis dataset of Google Maps reviews for the same 14 healthcare organizations in the HoPE-SA dataset. The newly constructed HEAR dataset contains 31,561 Arabic reviews extracted from Google Maps using outscraper.com, an online tool for scraping data. We named the dataset HEAR, which stands for Hospital Experiences Arabic Reviews.

We followed the same steps as for the pre-labeling of the HoPE-SA dataset in Section 3.1.2. The pre-labeling results are detailed in Table 4. As mentioned earlier, these labels required human verification before being used for the training of aspect-based sentiment analysis models.

The results indicated that 1992 instances were not classified into any of the initial categories. These instances were reviewed manually in the human verification stage and mapped to the three cases explained in Section 3.1.2.

Table 4. Pre-labeling results for the HEAR dataset using GPT-4 for aspect category polarity.

Polarity/Category	Negative	Positive	Neutral	Not Mentioned	Total
Medical Staff	7198	11,674	803	11,886	19,675
Appointments	5248	923	552	24,838	6723
Customer Service	10,781	11,365	809	8606	22,955
Emergency Services	3622	1077	479	26,383	5178
Pricing	3287	448	341	27,485	4076
Total	30,136	25,487	2,984	99,198	58,607

3.1.4. Dataset Annotation

To annotate the datasets, we recruited five annotators with Arabic as their mother tongue. We provided annotation guidelines to the annotators and assessed their understanding of the task by reviewing their annotations for ten reviews. The selection criteria for the annotators were as follows:

- Arabic native speaker;
- A minimum of 21 years old;
- A minimum of a bachelor's degree and/or professional certificates;
- Passing the initial annotation assessment.

The annotators were told to reach out if they had any doubts about the task. Each annotator was given a subset of the dataset, and each review/tweet was annotated by at least two annotators; in the case of disagreement, a third annotator was assigned for the review/tweet.

3.1.5. Dataset Cleaning

The datasets were reviewed manually to remove duplicate records and identify records with conflicting sentiments towards the same aspect category. After eliminating duplicate records and conflicting sentiments, the large number of reviews in the HEAR dataset was reduced to 25,156 reviews, whereas the number of instances in the HoPE-SA dataset

remained the same. For the HEAR dataset, we removed an extra 2152 reviews which were longer than 100 tokens, in order to reduce the computational complexity.

3.1.6. Exploratory Data Analysis (EDA)

A total of 17,155 mentioned aspect categories were identified in the HoPE-SA dataset. This dataset contains a total of 4889 positive tweets, 11,974 negative tweets, and 292 neutral tweets, with a total of 4984 tweets in the medical staff category, 1922 in the appointments category, 8607 in the customer service category, 750 in the pricing category, and 892 in the emergency services category. Table 5 provides further statistics on the HoPE-SA dataset. Figure 3 shows the percentage of tweets in each aspect category, and Figure 4 shows the percentage of tweets in each sentiment for each aspect category.

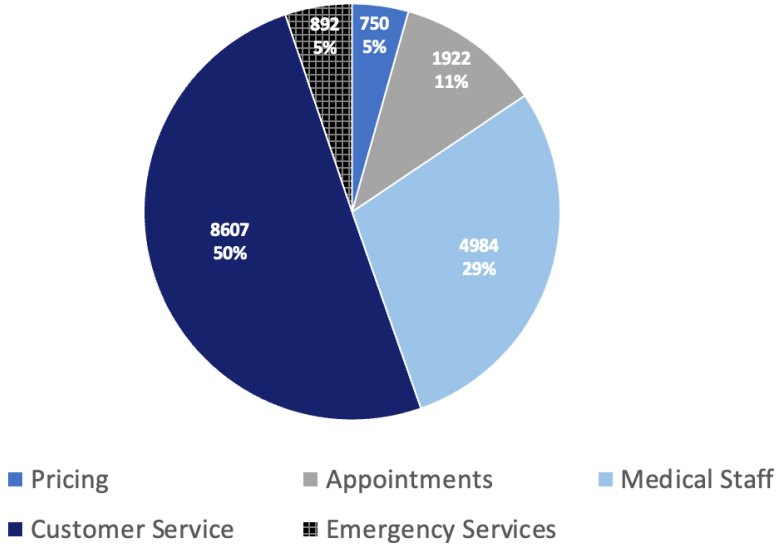

Figure 3. Visual representation of the percentages of each aspect category for the HoPE-SA dataset.

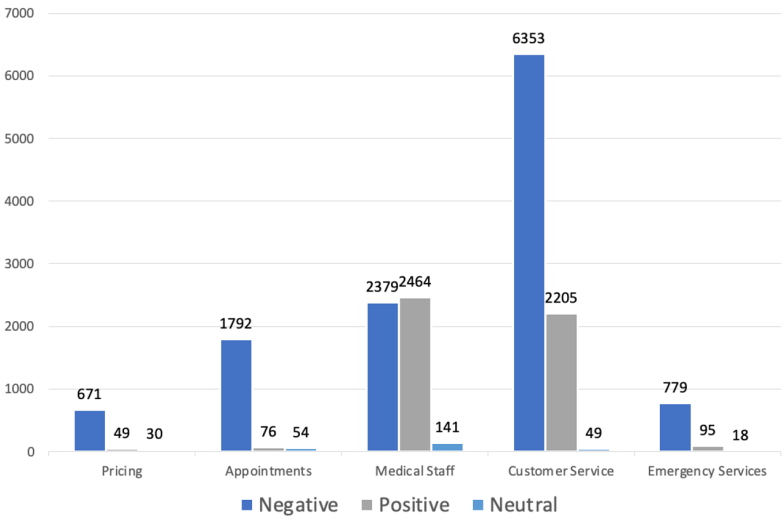

Figure 4. Visual representation of the percentages of each aspect category and polarity for the HoPE-SA dataset.

Table 5. Annotation results for the HoPE-SA dataset for aspect category polarity.

Polarity/Category	Negative	Positive	Neutral	Not Mentioned
Medical Staff	2379	2464	141	7416
Appointments	1792	76	54	10,478
Customer Service	6353	2205	49	3793
Emergency Services	779	95	18	11,508
Pricing	671	49	30	11,650

A total of 33,820 mentioned aspect categories were identified in the HEAR dataset. The dataset contains a total of 15,641 positive reviews, 17,133 negative reviews, and 1046 neutral reviews, with a total of 11,918 reviews in the medical staff category, 3139 in the appointments category, 14,718 in the customer service category, 1994 in the pricing category, and 2051 in the emergency services category. Table 6 provides more statistics for the HEAR dataset. Figure 5 shows the percentage of reviews in each aspect category, and Figure 6 shows the percentage of reviews in each sentiment for each aspect category.

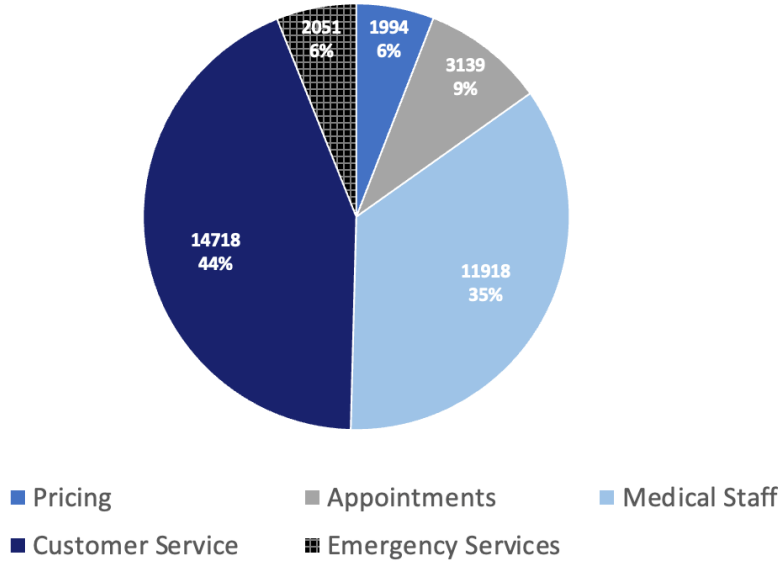

Figure 5. Visual representation of the percentages of each aspect category for the HEAR dataset.

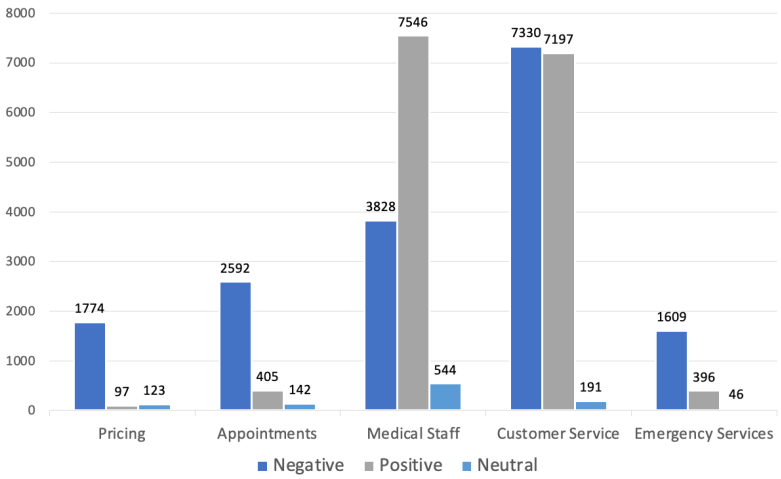

Figure 6. Visual representation of the percentages of each aspect category and polarity for the HEAR dataset.

Table 6. Annotation results for the HEAR dataset for aspect category polarity.

Polarity/Category	Negative	Positive	Neutral	Not Mentioned
Medical Staff	3828	7546	544	11,086
Appointments	2592	405	142	19,865
Customer Service	7330	7197	191	8286
Emergency Services	1609	396	46	20,953
Pricing	1774	97	123	21,010

3.2. Dataset Pre-Processing

Several pre-processing techniques were implemented to eliminate noise, unify the reviews/tweets, and improve the learning process. We utilized steps similar to those detailed by Almuhaideb et al. [2], who authored the HoPE-SA dataset. Regular expressions were employed to perform the following pre-processing steps:

- Removing links and mention tags;
- Removing punctuation;
- Removing underscores and hash symbols;
- Removing digits;
- Removing English words.

After pre-processing, we tokenized the data using BERT's WordPiece tokenizer, where we tokenized each review/tweet text to a length of 100 tokens. We added a special token [PAD] for reviews/tweets shorter than 100 tokens and, due to the associated computational complexity, we eliminated reviews longer than 100 tokens. We added the [CLS] token at the beginning and the [SEP] token at the end of each review/tweet text. In the sentence-pair models, we added another [SEP] token between the last token of the review/tweet text and the first token of the aspect category.

3.3. BERT-Based Models

This work proposes two BERT-based models; the first model solves the aspect category detection and aspect category polarity sub-tasks jointly, whereas the second model consists of two stages to solve the two sub-tasks separately.

3.3.1. Joint Model

This model jointly solves the aspect category detection and aspect category polarity sub-tasks as shown in Figure 7. The model uses BERT as a word embedding stacked with three well-known machine learning classifiers for aspect-based sentiment analysis for patient experience. The problem of aspect category detection and aspect category polarity is formulated as a sentence-pair classification problem using BERT. As suggested in the literature, formulating the problem as a sentence-pair classification problem using BERT yields better results, when compared with single-sentence classification. As illustrated in Figure 8, for a given text, *"Doctor Mohammed Shaker at Al Hammadi Hospital- Alsuwaidi (branch), has a lot of experience and is a very skilled artist"*, and aspect category, *"The medical Staff"*, the model will produce a sentiment polarity result. Thus, formulating the problems of aspect category detection and aspect category polarity as a sentence-pair classification problem allows the model to jointly solve these two sub-tasks. Figure 9 shows an example of the model input *"Excellent care and highly specialized excellent doctors, the hospital cleanliness is excellent and the location is close. The staff is very courteous in their dealings. We ask Allah to make them always rescuers for the patients with merciful hearts, O Lord"*, with the expected output.

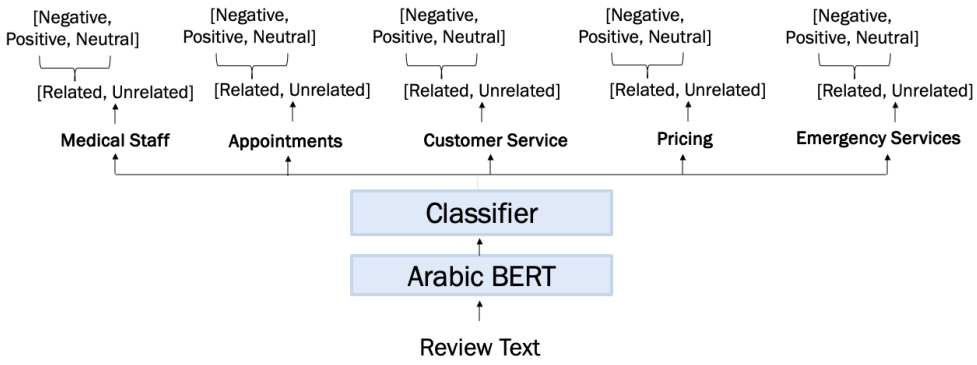

Figure 7. Illustration of the joint model for aspect-based sentiment analysis for patient experience.

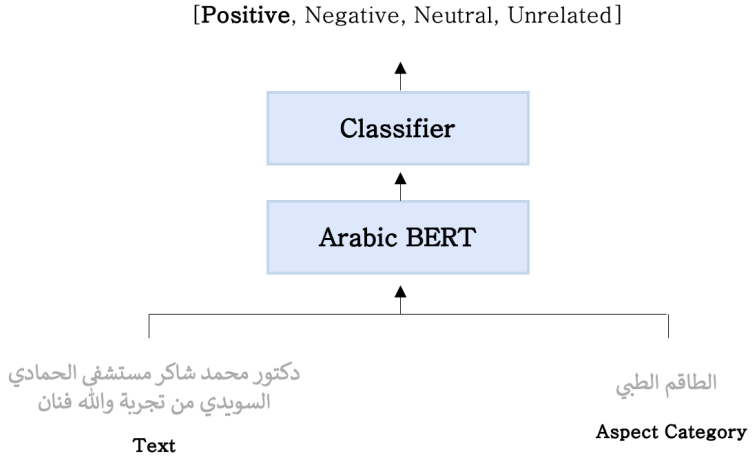

Figure 8. Illustration of aspect-based sentiment analysis for patient experience as a sentence-pair classification problem.

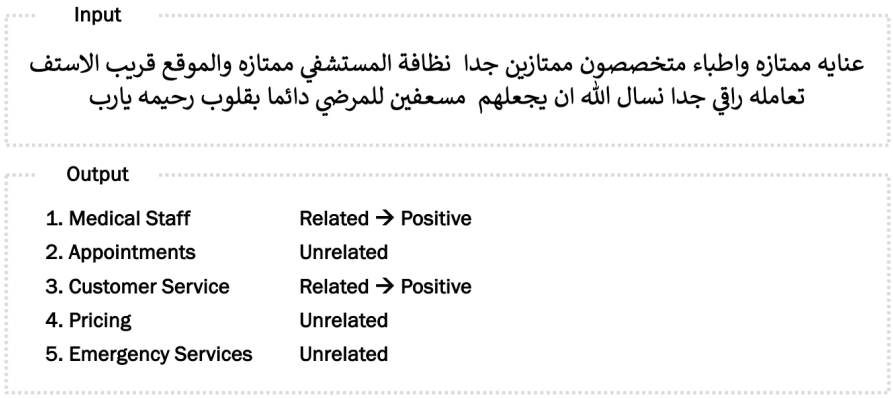

Figure 9. An example of aspect-based sentiment analysis input and expected output.

We fine-tuned five Arabic versions of BERT, namely, MARBERT [44], ArabicBERT [45], QARiB [46], CAMeLBERT [47], and AraBERT. The reasons for selecting these five versions of BERT are as follows:

- They are pre-trained specifically for the Arabic language.
- They are pre-trained fully or partially on dialectical Arabic corpus.

We used three well-known machine learning classifiers—namely, neural networks, SVM, and Random Forest (RF) [5]—in order to compare their performance in aspect-based sentiment analysis for patient experience.

The joint model is designed as a sentence-pair classification approach, which requires constructing the dataset in the sentence-pair format. We generated the possible combinations of review/tweet sentences and the aspect categories. Applying this construction method, the resulting dataset comprised pairs of review/tweet sentences and aspect categories with multi-class labels: '0', '1', '2', '3'. The label '0' means that the corresponding review/tweet sentence has a negative sentiment towards the paired aspect category, the label '1' means that the corresponding review/tweet sentence has a positive sentiment towards the paired aspect category, the label '2' means that the corresponding review/tweet sentence has a neutral sentiment towards the paired aspect category, and the label '3' means that the corresponding review/tweet sentence does not mention the paired aspect category. Table 7 provides an illustration of the constructed sentence-pair dataset.

Table 7. Illustration of constructed sentence-pair dataset for the joint model.

Review/Tweet Text (Sentence 1)	Aspect Category (Sentence 2)	Label
Nice receptionists but unqualified doctors	Medical Staff	0
Nice receptionists but unqualified doctors	Appointments	3
Nice receptionists but unqualified doctors	Customer Service	1
Nice receptionists but unqualified doctors	Emergency Services	3
Nice receptionists but unqualified doctors	Pricing	3

The training of the joint model consisted of two stages: the first stage was fine-tuning, and the second stage was classification. We split the dataset into 60% for training, 20% for validation, and 20% for testing. In the fine-tuning stage, the five Arabic versions of BERT were fine-tuned using a fully connected layer on the training and validation sets. The hidden layer size of the fully connected layer was 50, and the following parameters were set: the Adaptive Moment Estimation with Decoupled Weight Decay (AdamW) optimizer, a learning rate of 5×10^{-5}, a cross-entropy loss function, and two epochs of training with batch size equal to 32. In the classification stage, the embeddings produced from the fine-tuned BERT were fed to the three machine learning classifiers for training. We used the [CLS] vector as an input to the classification models. Due to limited computational power, we took a stratified sample representing 50% of the training set to train the classification models. After fine-tuning and training the models, we evaluated their performance on the test set.

3.3.2. Two-Stage Model

This model solves the aspect category detection and aspect category polarity as two sub-tasks using two sub-models. The first sub-model is a multi-label model that solves the aspect category detection problem using BERT followed by a fully connected layer. A certain threshold must be set to identify the detected aspect categories; in this way, the aspect categories below the threshold are labeled as unrelated, and the aspect categories above the threshold are considered related and fed to the second sub-model. The second sub-model is a multi-class model that solves the aspect category polarity problem as a sentence-pair classification problem using BERT followed by a fully connected layer.

Figure 10 provides an illustrative example that demonstrates the flow of the two-stage model, including the input and output of each sub-model, as outlined below:

- The first sub-model is a multi-label classification model that takes a review text as an input and produces five probabilities for the five aspect categories. In this example, the threshold is equal to 0.4 and, therefore, three of the five aspect categories are detected and fed to the second sub-model to determine their polarities. The other two aspect categories are not detected and so are labeled as unrelated.
- The second sub-model is a multi-class classification model that takes the review text as input along with each detected aspect category and determines the sentiment (i.e., as negative, positive, or neutral). The three sentiments in this example represent the polarities predicted by the second sub-model for the three detected aspect categories.

The value of the threshold hyper-parameter was set experimentally. Additionally, a comparative analysis was performed using the five Arabic versions of BERT discussed earlier.

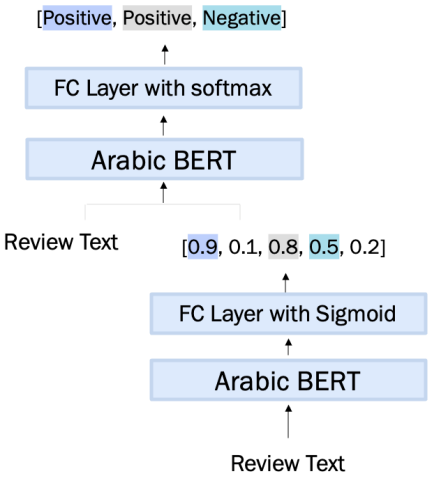

Figure 10. Illustration of the two-stage model for aspect-based sentiment analysis for patient experience.

The aspect category detection sub-model is a multi-label classification model that utilizes single-sentence BERT followed by a fully connected layer with a sigmoid activation function to detect one or more of the defined aspect categories. As the model is a single-sentence classification model, the dataset is constructed as the review/tweet text and its corresponding label in the form of multi one-hot encoding. The negative, positive, and neutral labels are transformed to the label '1' to represent the mentioned category, while the not mentioned label is transformed to the label '0'. Table 8 provides an illustration of the constructed single-sentence dataset, where the mentioned aspect categories (i.e., medical staff and customer service) are indicated by '1', while the other aspect categories (including pricing, appointments, and emergency services) are indicated by '0' to represent the not mentioned class. The training of the aspect category detection sub-model involved fine-tuning the five Arabic versions of BERT using a fully connected layer on the training and validation sets. We used the [CLS] vector as an input to the fully connected layer. The hidden layer size of the fully connected layer was 50, and the following parameters were set: the AdamW optimizer, a learning rate of 5×10^{-5}, a binary cross-entropy with logits loss function, and two epochs of training with batch size equal to 32. The training of the aspect category polarity sub-model involved fine-tuning the five Arabic versions of BERT using a fully connected layer on the training and validation sets. We used the [CLS] vector as an input to the fully connected layer. The hidden layer size of the fully connected layer was 50, and the following parameters were set: the AdamW optimizer, a learning rate of 5×10^{-5}, a cross-entropy loss function, and two epochs of training with batch size equal to 32.

Table 8. Illustration of constructed single-sentence dataset.

Review/Tweet Text	Label
Nice receptionists but unqualified doctors	[1, 0, 1, 0, 0]

The aspect category polarity sub-model is a multi-class classification model that utilizes sentence-pair BERT followed by a fully connected layer with a Softmax activation function to classify review/tweet text and aspect category pairs with respect to the corresponding sentiment. To construct the dataset for the aspect category polarity sub-model, we followed the same construction method as for the joint model. The only difference was that the labels for this model were '0', '1', and '2', where the label '0' means that the corresponding review/tweet sentence has a negative sentiment towards the paired

aspect category, the label '1' means that the corresponding review/tweet sentence has a positive sentiment towards the paired aspect category, and the label '2' means that the corresponding review/tweet sentence has a neutral sentiment towards the paired aspect category. We eliminated label '3', which represents the not mentioned aspects, as the task of detecting aspects was performed previously by the aspect category detection sub-model, while the objective of this sub-model was to determine the polarity with respect to each detected aspect. Table 9 provides an illustration of the constructed sentence-pair dataset.

Table 9. Illustration of constructed sentence-pair dataset for the aspect category polarity sub-model.

Review/Tweet Text (Sentence 1)	Aspect Category (Sentence 2)	Label
Nice receptionists but unqualified doctors	Medical Staff	0
Nice receptionists but unqualified doctors	Customer Service	1

We needed to evaluate the performance of the two-stage model in a similar manner to the joint model in order to effectively compare their performance. Therefore, during the evaluation, as illustrated in Figure 11, we fed the review/tweet sentences to the aspect category detection sub-model and produced the predicted outputs in the form of multi one-hot encoding. We transformed the predicted output into the form of a review/tweet sentence and aspect category pair for the detected aspect categories, and fed the transformed data to the aspect category polarity sub-model. The prediction of the aspect category polarity sub-model was then combined with the undetected aspect categories (i.e., those with label = '3').

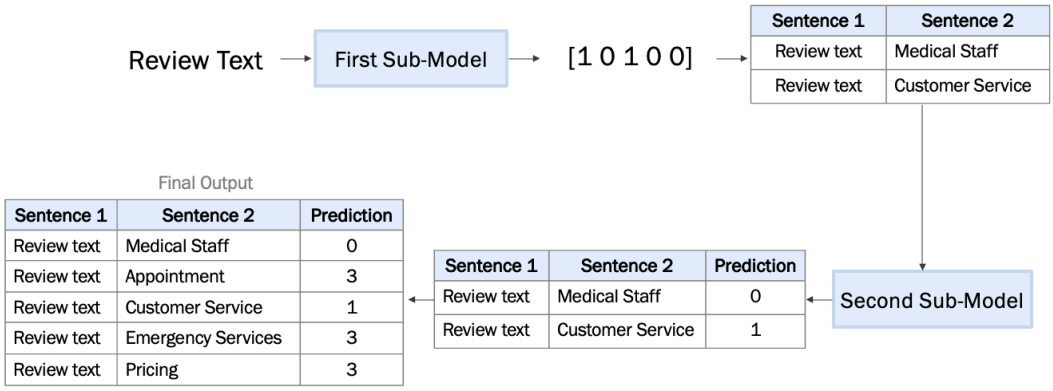

Figure 11. Illustration of evaluating the two-stage model.

3.4. Generative Model

This work also evaluated the performance of a generative model (i.e., GPT-4) for aspect category detection and aspect category polarity tasks. As mentioned earlier, GPT-4 was utilized for the pre-labeling of the datasets. These pre-labeled datasets were then subjected to human verification. However, these pre-labels were considered GPT-4 predictions and were used to evaluate the performance of GPT-4 in terms of aspect-based sentiment analysis for patient experience in few-shot settings.

4. Implementation Details and Evaluation Measures

BERT-based models were selected for our study due to their exceptional performance in various NLP tasks, including sentiment analysis. Their ability to effectively capture contextual information and the availability of pre-trained models tailored to the Arabic language made them particularly suitable for our research objectives. We also incorporated traditional machine learning classifiers as baseline models. These classifiers offer advantages such as simplicity and potentially quicker inference times in specific contexts, providing a valuable point of comparison for our findings. Random Forest (RF) [5] is an ensemble learning technique that combines multiple decision trees for improved prediction

accuracy. Each tree is trained on a randomly sampled subset of the data, and predictions are aggregated through averaging for regression or voting for classification. Based on the bagging principle, RF fosters diversity among trees, reducing variance and enhancing generalization performance. Support Vector Machine (SVM) is both a straightforward and effective method. Nguyen et al. [48] demonstrated that SVM can achieve high performance levels, even when compared with deep learning models and BERT variants. Therefore, we included SVM in our analysis of classical machine learning algorithms, as it is a widely used and proven competitive algorithm in prior research. Artificial Neural Networks (ANNs) [6] were used as a computational method to learn and predict intricate relationships within the dataset.

While alternative models, including Recurrent Neural Network (RNN)-based architectures and other transformer models such as T5 [49] and BART [24], were considered, they were ultimately excluded from our comparison, primarily based on concerns regarding computational cost and the risk of vanishing gradient issues associated with RNNs.

We utilized Google Colaboratory (Colab) [50] to build the training and testing environments, using a Colab Tensor Processing Unit (TPU) v2. All the implementations were carried out using Python 3.10.12 and PyTorch 2.5.0 [51] as a deep learning framework.

For the joint model, we conducted 15 different experiments using the five versions of Arabic BERT and the three classifiers (SVM, RF, and neural networks). The SVM utilized a linear kernel and a regularization parameter of 1.0. For RF, 100 trees were utilized with a Gini impurity criterion. As for the neural network, the size of the hidden layer was 50, and the Adaptive Moment Estimation (Adam) [52] optimizer was used with a learning rate of 0.001 and a cross-entropy loss function. For the two-stage model, we experimented nine values for the threshold hyper-parameter.

In terms of evaluation measures, the performance of the models was assessed using standard classification metrics, including accuracy, precision, recall, and F1-score. In addition to quantitative measures, confusion matrices and ROC curves were also used for performance comparison of the models. In multi-class classification tasks, where the model must distinguish between more than two classes, performance was evaluated using the average accuracy per class:

$$\text{Average Accuracy} = \frac{1}{I}\sum_{i=1}^{I}\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i},$$

where I is the number of classes, TP and TN represent the number of true positives and true negatives, respectively, and FP and FN represent the number of false positives and false negatives.

To assess the precision, recall, and F1-score across all classes, different averaging methods can be used. Macro-averaging computes the unweighted average of these metrics for each class:

$$\text{Precision (Macro)} = \frac{1}{I}\sum_{i=1}^{I}\frac{TP_i}{TP_i + FP_i},$$

$$\text{Recall (Macro)} = \frac{1}{I}\sum_{i=1}^{I}\frac{TP_i}{TP_i + FN_i},$$

$$\text{F1-score (Macro)} = \frac{2 \times \text{Precision (Macro)} \times \text{Recall (Macro)}}{\text{Precision (Macro)} + \text{Recall (Macro)}}.$$

Meanwhile, micro-averaging aggregates the *TP*, *FP*, *FN*, and *TN* counts across all classes before calculating the metrics:

$$\text{Precision (Micro)} = \frac{\sum_{i=1}^{I} TP_i}{\sum_{i=1}^{I} (TP_i + FP_i)},$$

$$\text{Recall (Micro)} = \frac{\sum_{i=1}^{I} TP_i}{\sum_{i=1}^{I} (TP_i + FN_i)},$$

$$\text{F1-score (Micro)} = \frac{2 \times \text{Precision (Micro)} \times \text{Recall (Micro)}}{\text{Precision (Micro)} + \text{Recall (Micro)}}.$$

Weighted-averaging accounts for class imbalances by weighting each class's metric by the number of instances in that class.

In multi-label classification, where each instance may belong to multiple classes, performance can be evaluated using the subset accuracy (exact match ratio), which considers a prediction correct only if all the predicted labels for an instance match the true labels:

$$\text{Subset Accuracy} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(L_{p_i} = L_{t_i}),$$

where n is the number of instances, L_{p_i} is the set of predicted labels, and L_{t_i} is the set of true labels for instance i.

The Hamming loss offers a less stringent measure by calculating the fraction of incorrect labels:

$$\text{Hamming Loss} = \frac{1}{nI} \sum_{i=1}^{n} \sum_{j=1}^{I} \mathbb{1}(L_{p_i}[j] \neq L_{t_i}[j]).$$

5. Results

5.1. Joint Model

The results obtained by the SVM, RF, and neural networks with the five versions of Arabic BERT are reported in Tables 10, 11 and 12, respectively, with the best scores shown in bold.

Table 10. Results of the joint model with SVM and the five versions of Arabic BERT.

BERT Model	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score
MARBERT	**92.14%**	**92.00%**	**92.14%**	**92.06%**
ArabicBERT	90.81%	90.72%	90.81%	90.76%
QARiB	91.27%	91.19%	91.27%	91.23%
CAMeLBERT	90.96%	90.88%	90.96%	90.92%
AraBERT	91.08%	90.86%	90.08%	90.95%

Table 11. Results of the joint model with RF and the five versions of Arabic BERT.

BERT Model	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score
MARBERT	**92.09%**	**91.96%**	**92.09%**	**92.01%**
ArabicBERT	91.46%	91.23%	91.46%	91.30%
QARiB	91.69%	91.54%	91.69%	91.60%
CAMeLBERT	91.73%	91.57%	91.73%	91.62%
AraBERT	91.39%	91.19%	91.39%	91.25%

Table 12. Results of the joint model with the neural networks and the five versions of Arabic BERT.

BERT Model	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score
MARBERT	**91.47%**	**91.44%**	**91.47%**	**91.45%**
ArabicBERT	90.13%	90.10%	90.13%	90.10%
QARiB	90.91%	90.76%	90.91%	90.82%
CAMeLBERT	90.43%	90.41%	90.43%	90.42%
AraBERT	90.17%	90.04%	90.17%	90.10%

Figure 12 shows the ROC curves of the three classifiers with the MARBERT model, and Figure 13 shows the confusion matrices for the three classifiers with the MARBERT model.

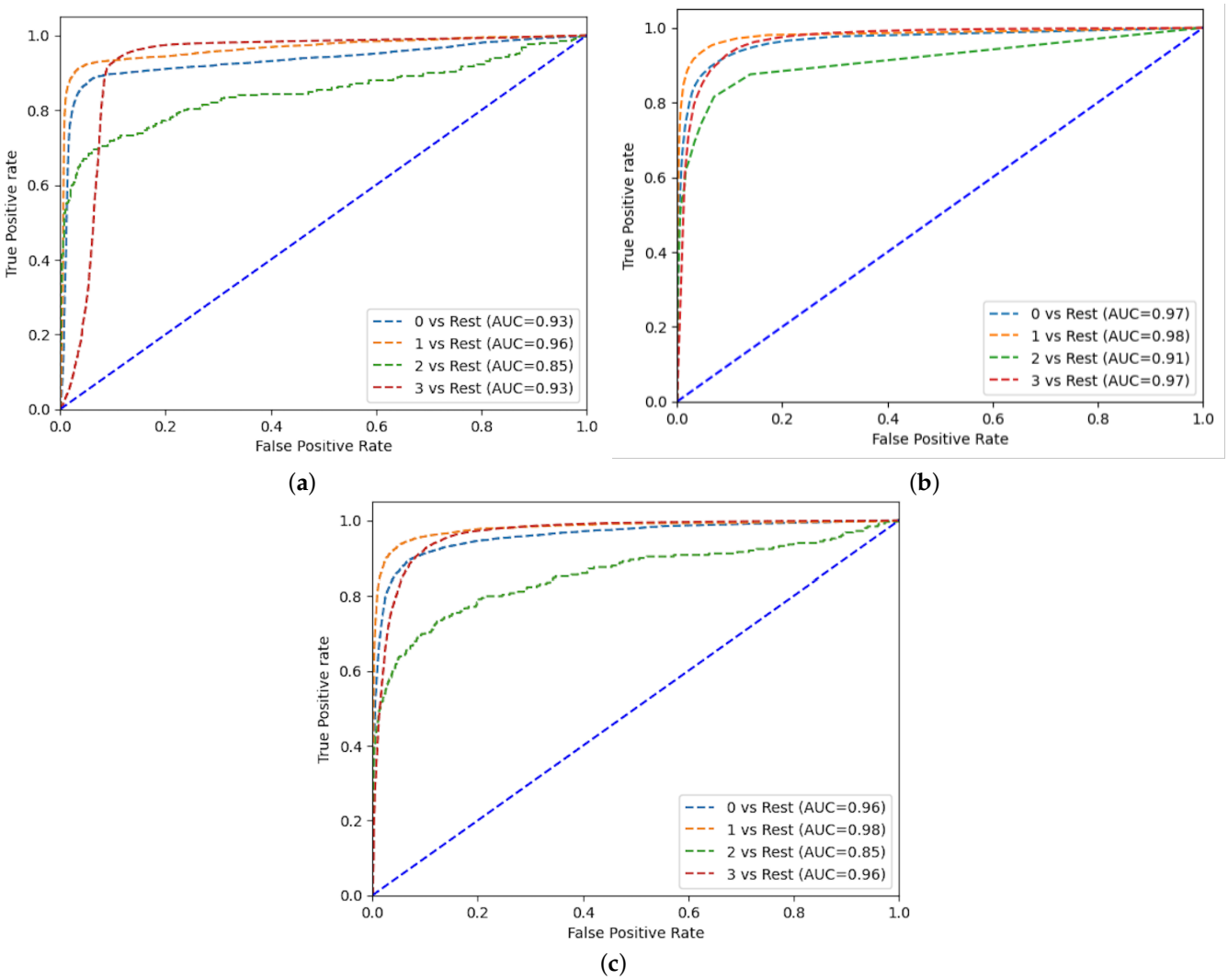

Figure 12. ROC curves for the three classifiers of the joint model with the MARBERT model: (**a**) SVM; (**b**) RF; (**c**) neural network.

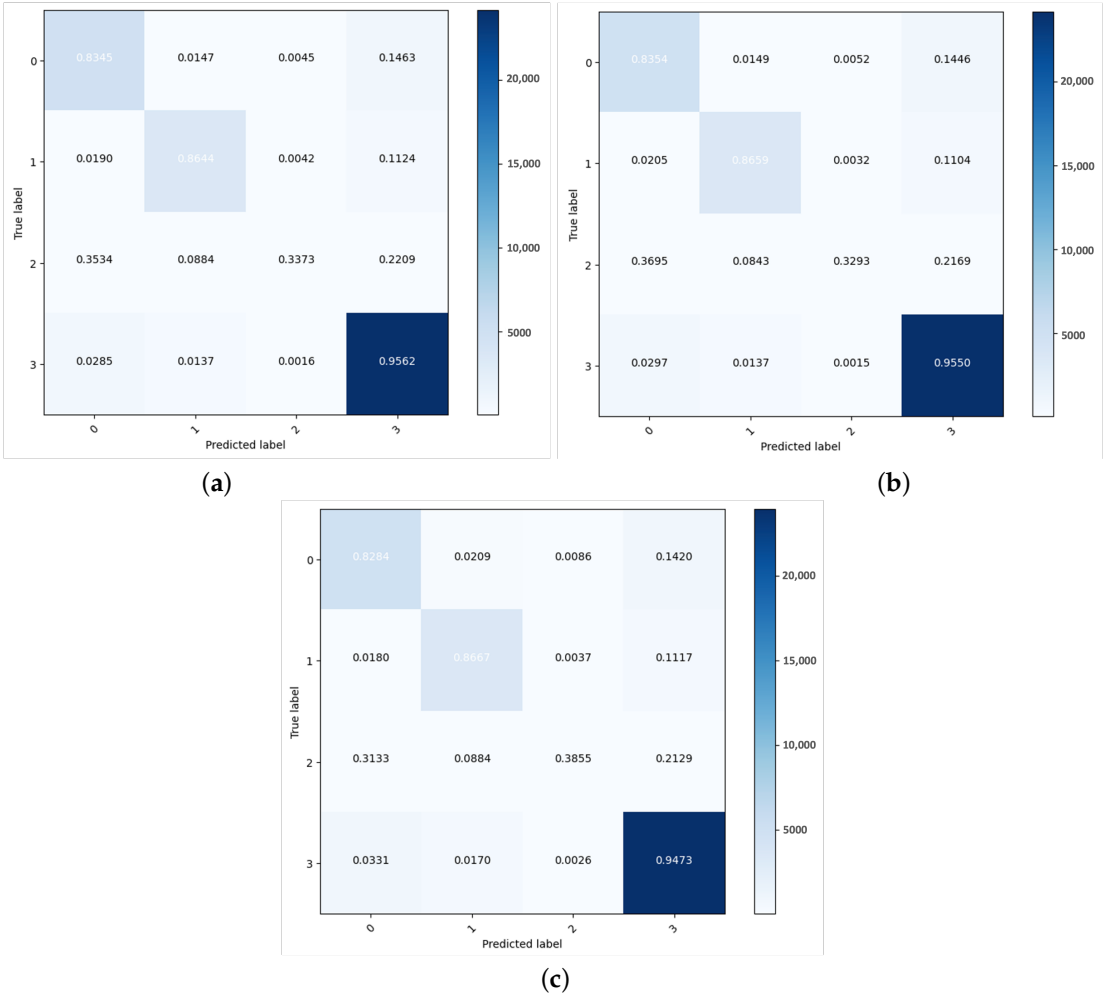

Figure 13. Confusion matrices for the three classifiers of the joint model with MARBERT model: (**a**) SVM; (**b**) RF; (**c**) neural network.

5.2. Two-Stage Model

The results obtained by the two-stage model using MARBERT with different values of the threshold parameter are reported in Table 13, showing the best scores in bold.

Table 13. Results of two-stage model using MARBERT with different values of the threshold hyper-parameter.

Threshold Value	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score
0.1	76.03%	**76.68%**	76.03%	74.22%
0.2	80.07%	76.72%	80.07%	77.07%
0.3	80.07%	76.72%	80.07%	77.07%
0.4	81.87%	76.26%	81.87%	78.17%
0.5	82.28%	76.00%	82.28%	**78.37%**
0.6	**82.46%**	75.59%	**82.46%**	78.36%
0.7	82.38%	74.85%	82.38%	78.09%
0.8	81.89%	73.46%	81.89%	77.29%
0.9	79.83%	70.08%	79.83%	74.64%

The results obtained by the two-stage model with the five versions of Arabic BERT and the best threshold value based on weighted F1-score are reported in Table 14, with the best scores shown in bold. Figure 14 shows the confusion matrix for the two-stage model with the QARiB model.

Table 14. Results for the two-stage model with the five versions of Arabic BERT.

BERT Model	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score
MARBERT	82.28%	76.00%	82.28%	78.37%
ArabicBERT	82.12%	75.64%	82.12%	78.15%
QARiB	**82.37%**	**75.93%**	**82.37%**	**78.40%**
CAMeLBERT	82.23%	75.74%	82.23%	78.26%
AraBERT	68.85%	67.08%	68.85%	67.70%

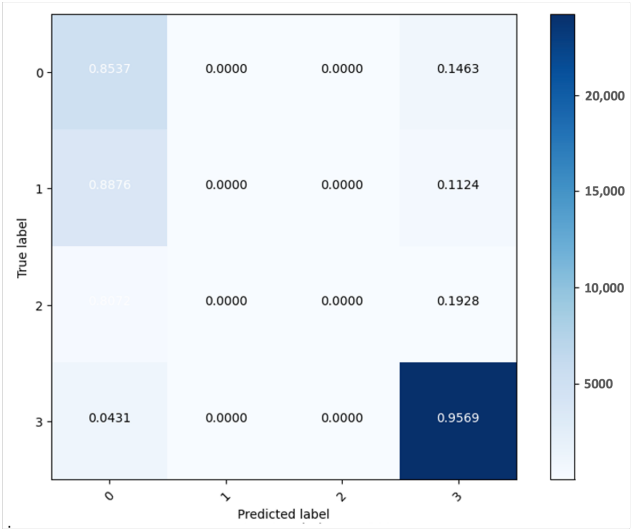

Figure 14. Confusion matrix for the two-stage model with the QARiB model.

5.2.1. Aspect Category Detection Sub-Model

This subsection provides the results of the aspect category detection sub-model, where we evaluate the performance of the sub-model in the aspect category detection sub-task. Table 15 provides the results for the aspect category detection sub-model with the five versions of Arabic BERT and the best threshold value based on the weighted F1-score, with the best scores shown in bold.

Table 15. Results for the aspect category detection sub-model with the five versions of Arabic BERT.

BERT Model	Hamming Loss	Weighted Precision	Weighted Recall	Weighted F1-Score
MARBERT	**0.290**	**48.32%**	**48.30%**	**48.30%**
ArabicBERT	0.293	48.01%	47.08%	47.53%
QARiB	0.295	47.49%	46.89%	47.18%
CAMeLBERT	0.292	48.14%	47.19%	47.65%
AraBERT	0.292	47.85%	46.47%	47.14%

5.2.2. Aspect Category Polarity Sub-Model

This subsection provides the results for the aspect category polarity sub-model, where we evaluate the performance of the sub-model in the aspect category polarity sub-task. Table 16 provides the results of the aspect category polarity sub-model with the five versions of Arabic BERT, with the best scores shown in bold.

Table 16. Results for the aspect category polarity sub-model with the five versions of Arabic BERT.

BERT Model	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score
MARBERT	49.72%	49.04%	49.72%	49.38%
ArabicBERT	**55.86%**	49.24%	**55.86%**	45.81%
QARiB	49.05%	48.47%	49.05%	48.76%
CAMeLBERT	50.40%	**49.64%**	50.40%	**50.01%**
AraBERT	50.19%	49.60%	50.19%	49.89%

5.3. Generative Model

Regarding the generative model, we evaluated the performance of GPT-4 in the aspect category detection and aspect category polarity tasks. The results obtained by GPT-4 in few-shot settings are reported in Table 17. Figure 15 shows the confusion matrix for GPT-4.

Table 17. Results for GPT-4 for aspect category detection and polarity.

Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score
88.88%	90.66%	88.88%	89.33%

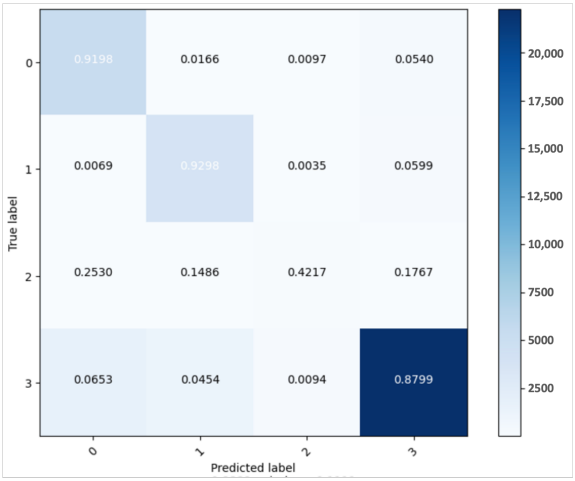

Figure 15. Confusion matrix for GPT-4.

6. Discussion

The first contribution of this work is the introduction of a Google Maps-based dataset; namely, an Arabic patient experience review dataset related to healthcare providers within Saudi Arabia, which is annotated with respect to several healthcare aspects and different sentiment polarities. In addition, we annotated the HoPE-SA dataset for the aspect category detection and aspect category polarity tasks. One of the objectives of this work was to analyze patient experiences in Saudi Arabia concerning several healthcare aspects. Based on the annotation of both datasets and from the statistical analysis performed on these datasets, the most-discussed aspect was customer service, followed by medical staff. In general, there was a higher rate of negative sentiments towards healthcare aspects than other sentiments.

The second contribution of this work is a comparative study related to the use of fine-tuned BERT as word embedding with machine learning classifiers—namely, SVM, RF, and neural networks—for the tasks of aspect category detection and polarity. As illustrated in Figure 16, the experimental results revealed that the RF and SVM classifiers outperformed the neural networks, which can be considered a baseline model, with all different versions of Arabic BERT. Moreover, the comparison between the five Arabic versions of BERT in the joint model indicated that the MARBERT-based models had superior performance among the five models with all three classifiers.

The experimental results comparing the joint model's performance with the two-stage model indicated that the joint model surpassed the two-stage model in the aspect category detection and polarity tasks with a large margin of enhancement. We concluded that not only did the sentence-pair classification problem formulation contribute to its superior performance but also the joint model was fine-tuned using a larger amount of data, allowing it to learn a better language representation. The larger dataset utilized for fine-tuning significantly contributed to the superior performance of the joint model over the two-stage model. This underscores the critical role of the dataset size in model training and evaluation, as larger datasets typically provide more diverse and representative examples, thereby enhancing the model's learning capacity and its ability to generalize to unseen data. In the two-stage model, the second sub-model was formulated as sentence-pair classification; however, it was fine-tuned on a smaller subset of the dataset, as it is only an aspect category polarity model and, thus, we eliminated the "not mentioned" class instances.

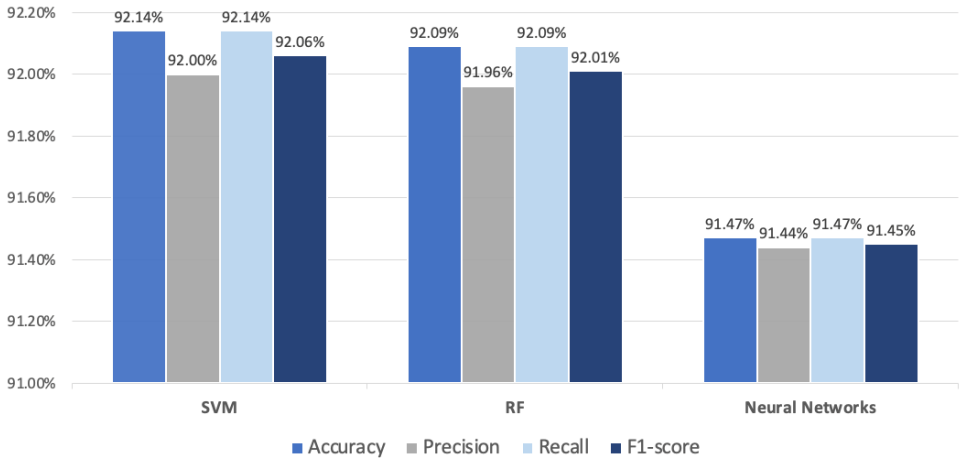

Figure 16. Visual representation of the joint model results using MARBERT model and the three classifiers.

Experimental results comparing the performance of the generative model (i.e., GPT-4) with that of the BERT-based models revealed that the joint BERT-based model outperformed GPT-4 in the tasks of aspect category detection and aspect category polarity, although with a relatively small margin of enhancement. It is worth mentioning that the GPT-4 results were obtained in few-shot training settings. Therefore, its performance could be improved through the use of more training shots.

This study has several potential limitations. First, the training of the joint model was conducted using a two-stage approach, rather than an end-to-end approach, which might yield better results. Second, the reviews collected from Twitter and Google Maps may not represent the entire patient population. Certain demographics might be under-represented, which could lead to biased sentiment results. Additionally, the findings from this specific study may not be generalizable to other regions or healthcare contexts, limiting the applicability of the results. The third limitation involves the challenge associated with accurately identifying the relevant features and aspects from text data. Incomplete or incorrect extraction can result in misleading sentiment classifications, which is especially critical in healthcare, where nuanced opinions impact patient care and decision-making. We recognize that the effectiveness of aspect-based sentiment analysis depends on this accuracy, and that overlooking or misclassifying key elements may lead to erroneous results. This highlights the need for ongoing advancements in methodologies for aspect detection and feature extraction. Finally, we recognize that patient sentiment is not static, and can change over time due to various factors such as evolving treatment experiences and shifts in healthcare quality. Consequently, the analyzed reviews may capture sentiments that reflect a specific moment, rather than an ongoing patient experience. Acknowledging this temporal dimension is essential for a comprehensive understanding of sentiment dynam-

ics in healthcare, and suggests opportunities for future research to explore longitudinal approaches to sentiment analysis.

Another challenge is that generic sentiment analysis, as evidenced in some existing research, often fails to capture the specificity and complexity of patient sentiments regarding particular aspects of healthcare. Through focusing on aspect-based sentiment analysis, we can better understand and represent the varied experiences and opinions of patients, leading to more accurate and actionable insights in the healthcare domain.

The findings of this study have significant implications for improving healthcare services. By linking specific feedback aspects (e.g., concerns about long wait times) to actionable solutions (e.g., optimizing scheduling systems and increasing staffing during peak hours), healthcare providers can enhance patient satisfaction and outcomes. Furthermore, incorporating patient feedback fosters a patient-centered approach, allowing providers to better meet the needs and preferences of individuals. Additionally, the study's results can guide policy recommendations to address recurring issues and revise care protocols as necessary. These findings can assist healthcare providers in the Kingdom in prioritizing their improvement plans based on patient opinions, ultimately leading to significant enhancements in overall patient satisfaction.

7. Conclusions

Aspect-based sentiment analysis is a fine-grained analysis with the aim of extracting the aspects discussed in a text and their polarities. In the healthcare domain, utilizing such an advanced analysis provides a comprehensive view of the strengths and weaknesses of the provided healthcare service with respect to various aspects, which enables the healthcare provider to make targeted improvements. This study aimed to bridge the gap related to Arabic aspect-based sentiment analysis in the healthcare domain, annotate the HoPE-SA dataset, and introduce an annotated Arabic patient experience reviews dataset for aspect category detection and polarity tasks. The newly constructed HEAR dataset contains Arabic reviews related to patient experiences derived from Google Maps. In this work, we annotated the HoPE-SA and HEAR datasets with five aspect categories—namely, medical staff, appointments, customer service, emergency services, and pricing—along with four sentiment classes: negative, positive, neutral, and not mentioned.

In this work, a comparative study of different BERT-based models was conducted to investigate the performance of BERT in word embedding with machine learning classifiers, namely, neural networks, SVM, and RF. We fine-tuned five different Arabic pre-trained BERT models: MARBERT, ArabicBERT, AraBERT, QARiB, and CAMeLBERT. In addition, we evaluated the performance of GPT-4 using ChatGPT by OpenAI in terms of aspect-based sentiment analysis for patient experience. The experimental results demonstrate the superiority of the MARBERT model, when used with the SVM and RF classifiers, in capturing aspect-based sentiment.

The results show that the joint BERT-based model outperformed the two-stage model and GPT-4 in the aspect-based sentiment analysis task, with MARBERT achieving the highest accuracy at 92.14% and a weighted F1-score of 92.06% using the SVM classifier. The two-stage model with QARiB also showed strong performance, with an F1-score of 78.40%.

As future work, we may consider longer sentences using techniques such as sliding windows. We could also evaluate the proposed approaches in other aspect-based sentiment analysis tasks. In addition, it would be interesting to explore specifying the levels of sentiments or emotions instead of defining the sentiments as simply negative, positive, or neutral.

8. Ethical Considerations

The datasets utilized in this research were collected from public sources (i.e., Twitter and Google Maps) and only included the review text without any personal identifiers (e.g., account names). When utilizing data from social media platforms, ethical considerations are paramount. Users understandably express concerns regarding the privacy and confi-

dentiality of their online information. Notably, data collected from the web for social good or public health purposes are often viewed as more acceptable in social media research by users [53,54]. While we recognize these concerns, it is important to note that all information used in this study was obtained for academic research purposes and was limited to publicly available posts that users chose not to keep private. The processing of these data in this research is in accordance with the Personal Data Protection Law [55], which applies to any processing of personal data related to individuals that takes place in the Kingdom. The dataset annotation process was conducted with sensitive attention to cultural and linguistic nuances. The annotators were trained to follow specific guidelines, ensuring the consistent and respectful interpretation of sentiments.

Author Contributions: Conceptualization, S.A. (Seba AlNasser) and S.A. (Sarab AlMuhaideb); methodology, S.A. (Seba AlNasser) and S.A. (Sarab AlMuhaideb); software, S.A. (Seba AlNasser); validation, S.A. (Seba AlNasser); formal analysis, S.A. (Seba AlNasser) and S.A. (Sarab AlMuhaideb); writing—original draft preparation, S.A. (Seba AlNasser); writing—review and editing, S.A. (Sarab AlMuhaideb); supervision, S.A. (Sarab AlMuhaideb). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The annotated version of the HoPE-SA dataset for aspect-based sentiment analysis and the newly constructed HEAR dataset are publicly available at: https://github.com/Sebalnasser/Arabic-Aspect-Based-Sentiment-Analysis-for-Patient-Experience (accessed on 13 November 2024).

Acknowledgments: The authors thank the anonymous reviewers for their constructive comments.

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

References

1. Weinman, J. Doctor–Patient Interaction: Psychosocial Aspects. In *International Encyclopedia of the Social & Behavioral Sciences*; Smelser, N.J., Baltes, P.B., Eds.; Pergamon: Oxford, UK, 2001; pp. 3816–3821. [CrossRef]
2. AlMuhaideb, S.; AlNegheimish, Y.; AlOmar, T.; AlSabti, R.; AlKathery, M.; AlOlyyan, G. Analyzing Arabic Twitter-Based Patient Experience Sentiments Using Multi-Dialect Arabic Bidirectional Encoder Representations from Transformers. *Comput. Mater. Contin.* **2023**, *76*, 195–220. [CrossRef]
3. Pontiki, M.; Galanis, D.; Pavlopoulos, J.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, 23–24 August 2014; pp. 27–35. [CrossRef]
4. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
5. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
6. McCulloch, W.S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **1943**, *5*, 115–133. [CrossRef]
7. Alassaf, M.; Qamar, A.M. Aspect-based sentiment analysis of Arabic tweets in the education sector using a hybrid feature selection method. In Proceedings of the 2020 14th International Conference on Innovations in Information Technology (IIT), Al Ain, United Arab Emirates, 17–18 November 2020; pp. 178–185. [CrossRef]
8. AlMasaud, A.; Al-Baity, H.H. AraMAMS: Arabic Multi-Aspect, Multi-Sentiment Restaurants Reviews Corpus for Aspect-Based Sentiment Analysis. *Sustainability* **2023**, *15*, 12268. [CrossRef]
9. Robbins, H.; Monro, S. A Stochastic Approximation Method. *Ann. Math. Stat.* **1951**, *22*, 400–407. [CrossRef]
10. Al-Smadi, M.; Talafha, B.; Al-Ayyoub, M.; Jararweh, Y. Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews. *Int. J. Mach. Learn. Cybern.* **2019**, *10*, 2163–2175. [CrossRef]
11. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
12. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**. [CrossRef]
13. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
14. Lafferty, J.D.; McCallum, A.; Pereira, F.C.N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01), Williamstown, MA, USA, 28 June 28–1 July 2001; pp. 282–289.
15. Kuppusamy, M.; Selvaraj, A. A novel hybrid deep learning model for aspect based sentiment analysis. *Concurr. Comput. Pract. Exp.* **2023**, *35*, e7538. [CrossRef]

16. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **1980**, *36*, 193–202. [CrossRef] [PubMed]
17. Han, Y.; Liu, M.; Jing, W. Aspect-level drug reviews sentiment analysis based on double BiGRU and knowledge transfer. *IEEE Access* **2020**, *8*, 21314–21325. [CrossRef]
18. Sivakumar, M.; Uyyala, S.R. Aspect-based sentiment analysis of mobile phone reviews using LSTM and fuzzy logic. *Int. J. Data Sci. Anal.* **2021**, *12*, 355–367. [CrossRef]
19. Gao, Z.; Li, Z.; Luo, J.; Li, X. Short text aspect-based sentiment analysis based on CNN+ BiGRU. *Appl. Sci.* **2022**, *12*, 2707. [CrossRef]
20. Al-Dabet, S.; Tedmori, S.; Mohammad, A.S. Enhancing Arabic aspect-based sentiment analysis using deep learning models. *Comput. Speech Lang.* **2021**, *69*, 101224. [CrossRef]
21. Gonnet, P.; Deselaers, T. Indylstms: Independently recurrent LSTMs. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 3352–3356. [CrossRef]
22. Apostol, E.S.; Pisică, A.G.; Truică, C.O. ATESA-BÆRT: A Heterogeneous Ensemble Learning Model for Aspect-Based Sentiment Analysis. *arXiv* **2023**, arXiv:2307.15920.
23. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
24. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 7871–7880. [CrossRef]
25. Rani, S.; Jain, A. Aspect-based sentiment analysis of drug reviews using multi-task learning based dual BiLSTM model. *Multimed. Tools Appl.* **2023**, 83, 22473–22501. [CrossRef]
26. Chouikhi, H.; Alsuhaibani, M.; Jarray, F. BERT-Based Joint Model for Aspect Term Extraction and Aspect Polarity Detection in Arabic Text. *Electronics* **2023**, *12*, 515. [CrossRef]
27. Li, X.; Bing, L.; Zhang, W.; Lam, W. Exploiting BERT for End-to-End Aspect-based Sentiment Analysis. In Proceedings of the 5th Workshop on Noisy User-Generated Text (W-NUT 2019), Hong Kong, China, 4 November 2019; pp. 34–41. [CrossRef]
28. Gao, Y.; Glowacka, D. Deep Gate Recurrent Neural Network. In Proceedings of the 8th Asian Conference on Machine Learning, Hamilton, New Zealand, 16–18 November 2016; Durrant, R.J., Kim, K.E., Eds.; The University of Waikato: Hamilton, New Zealand, 2016; Volume 63, pp. 350–365.
29. Abdelgwad, M.M.; Soliman, T.H.A.; Taloba, A.I. Arabic aspect sentiment polarity classification using BERT. *J. Big Data* **2022**, *9*, 115. [CrossRef]
30. Fadel, A.S.; Saleh, M.E.; Abulnaja, O.A. Arabic aspect extraction based on stacked contextualized embedding with deep learning. *IEEE Access* **2022**, *10*, 30526–30535. [CrossRef]
31. Antoun, W.; Baly, F.; Hajj, H. AraBERT: Transformer-based Model for Arabic Language Understanding. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, Marseille, France, 11–16 May 2020; pp. 9–15.
32. Akbik, A.; Blythe, D.; Vollgraf, R. Contextual String Embeddings for Sequence Labeling. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 1638–1649.
33. Hopfield, J.J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA* **1982**, *79*, 2554–2558. [CrossRef] [PubMed]
34. Chang, Y.C.; Ku, C.H.; Nguyen, D.D.L. Predicting aspect-based sentiment using deep learning and information visualization: The impact of COVID-19 on the airline industry. *Inf. Manag.* **2022**, *59*, 103587. [CrossRef]
35. Hoang, M.; Bihorac, O.A.; Rouces, J. Aspect-Based Sentiment Analysis using BERT. In Proceedings of the 22nd Nordic Conference on Computational Linguistics, Turku, Finland, 30 September–2 October 2019; pp. 187–196.
36. Shim, H.; Lowet, D.; Luca, S.; Vanrumste, B. Lets: A label-efficient training scheme for aspect-based sentiment analysis by using a pre-trained language model. *IEEE Access* **2021**, *9*, 115563–115578. [CrossRef]
37. Sun, C.; Huang, L.; Qiu, X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. *arXiv* **2019**. [CrossRef]
38. Li, X.; Fu, X.; Xu, G.; Yang, Y.; Wang, J.; Jin, L.; Liu, Q.; Xiang, T. Enhancing BERT representation with context-aware embedding for aspect-based sentiment analysis. *IEEE Access* **2020**, *8*, 46868–46876. [CrossRef]
39. Al-Jarrah, I.; Mustafa, A.M.; Najadat, H. Aspect-Based Sentiment Analysis for Arabic Food Delivery Reviews. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2023**, *22*, 1–18. [CrossRef]
40. Lan, W.; Chen, Y.; Xu, W.; Ritter, A. An empirical study of pre-trained transformers for Arabic information extraction. *arXiv* **2020**. [CrossRef]
41. Bensoltane, R.; Zaki, T. Towards Arabic aspect-based sentiment analysis: A transfer learning-based approach. *Soc. Netw. Anal. Min.* **2022**, *12*, 7. [CrossRef]
42. OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Leoni Aleman, F.; Almeida, D.; Altenschmidt, J.; Altman, S.; et al. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774. [CrossRef]
43. Kheiri, K.; Karimi, H. SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning. *arXiv* **2023**, arXiv:2307.10234.

44. Abdul-Mageed, M.; Elmadany, A.; Nagoudi, E.M.B. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 1–6 August 2021; pp. 7088–7105. [CrossRef]
45. Safaya, A.; Abdullatif, M.; Yuret, D. KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona (online), 12–13 December 2020; pp. 2054–2059. [CrossRef]
46. Abdelali, A.; Hassan, S.; Mubarak, H.; Darwish, K.; Samih, Y. Pre-Training BERT on Arabic Tweets: Practical Considerations. *arXiv* **2021**, arXiv:2102.10684.
47. Inoue, G.; Alhafni, B.; Baimukan, N.; Bouamor, H.; Habash, N. The interplay of variant, size, and task type in Arabic pre-trained language models. *arXiv* **2021**. [CrossRef]
48. Nguyen, Q.T.; Nguyen, T.L.; Luong, N.H.; Ngo, Q.H. Fine-tuning bert for sentiment analysis of vietnamese reviews. In Proceedings of the 2020 7th NAFOSTED Conference on Information and Computer Science (NICS), Ho Chi Minh City, Vietnam, 26–27 November 2020; pp. 302–307.
49. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
50. Bisong, E. Google Colaboratory. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*; Bisong, E., Ed.; Apress: Berkeley, CA, USA, 2019; pp. 59–64. [CrossRef]
51. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in PyTorch. In Proceedings of the Neural Information Processing Systems (NIPS) Autodiff Workshop, Long Beach, CA, USA, 4–9 December 2017.
52. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015. [CrossRef]
53. Conway, M. Ethical issues in using Twitter for public health surveillance and research: Developing a taxonomy of ethical concepts from the research literature. *J. Med. Internet Res.* **2014**, *16*, e290. [CrossRef]
54. Golder, S.; Ahmed, S.; Norman, G.; Booth, A. Attitudes toward the ethics of research using social media: A systematic review. *J. Med. Internet Res.* **2017**, *19*, e195. [CrossRef]
55. Personal Data Protection Law. Royal Decree M/19 of 9/2/1443H (16 September 2021); Cabinet Resolution No. 98 of 7/2/1443H (14 September 2021). Available online: https://sdaia.gov.sa/en/SDAIA/about/Documents/Personal%20Data%20English%20V2-23April2023-%20Reviewed-.pdf (accessed on 4 November 2024).