computers

MDPI

Article

MTL-AraBERT: An Enhanced Multi-Task Learning Model for Arabic Aspect-Based Sentiment Analysis

Arwa Fadel [1,2,*], Mostafa Saleh [1], Reda Salama [1] and Osama Abulnaja [1]

[1] Computer Science Department, Faculty of Computing and Information Technology, King Abdulaziz University (KAU), Jeddah 21589, Saudi Arabia; msherbini@kau.edu.sa (M.S.); abulnaja@kau.edu.sa (O.A.)
[2] Computer Science Department, Faculty of Computer Sciences and Engineering, Hodeidah University, Hodeidah 207416, Yemen
* Correspondence: afadl@stu.kau.edu.sa

Abstract: Aspect-based sentiment analysis (ABSA) is a fine-grained type of sentiment analysis; it works on an aspect level. It mainly focuses on extracting aspect terms from text or reviews, categorizing the aspect terms, and classifying the sentiment polarities toward each aspect term and aspect category. Aspect term extraction (ATE) and aspect category detection (ACD) are interdependent and closely associated tasks. However, the majority of the current literature on Arabic aspect-based sentiment analysis (ABSA) deals with these tasks individually, assumes that aspect terms are already identified, or employs a pipeline model. Pipeline solutions employ single models for each task, where the output of the ATE model is utilized as the input for the ACD model. This sequential process can lead to the propagation of errors across different stages, as the performance of the ACD model is influenced by any errors produced by the ATE model. Therefore, the primary objective of this study was to investigate a multi-task learning approach based on transfer learning and transformers. We propose a multi-task learning model (MTL) that utilizes the pre-trained language model (AraBERT), namely, the MTL-AraBERT model, for extracting Arabic aspect terms and aspect categories simultaneously. Specifically, we focused on training a single model that simultaneously and jointly addressed both subtasks. Moreover, this paper also proposes a model integrating AraBERT, single pair classification, and BiLSTM/BiGRU that can be applied to aspect term polarity classification (APC) and aspect category polarity classification (ACPC). All proposed models were evaluated using the SemEval-2016 annotated dataset for the Arabic hotel dataset. The experiment results of the MTL model demonstrate that the proposed models achieved comparable or better performance than state-of-the-art works (F1-scores of 80.32% for the ATE and 68.21% for the ACD). The proposed SPC-BERT model demonstrated high accuracy, reaching 89.02% and 89.36 for APC and ACPC, respectively. These improvements hold significant potential for future research in Arabic ABSA.

Keywords: aspect-based sentiment analysis; AraBERT; deep learning; multi-task learning

Citation: Fadel, A.; Saleh, M.; Salama, R.; Abulnaja, O. MTL-AraBERT: An Enhanced Multi-Task Learning Model for Arabic Aspect-Based Sentiment Analysis. *Computers* **2024**, *13*, 98. https://doi.org/10.3390/computers13040098

Academic Editors: Yorghos Voutos, Akrivi Krouska, Christos Troussas, Phivos Mylonas and Cleo Sgouropoulou

Received: 9 March 2024
Revised: 8 April 2024
Accepted: 10 April 2024
Published: 15 April 2024

1. Introduction

Sentiment analysis (SA) has gained considerable scholarly interest in the NLP community in recent years. Early research on SA concentrated on document-level and sentence-level classifications, which evaluate an overall document (e.g., a movie review) or sentences as positive, negative, or neutral. However, classifying opinionated text as atomic units is insufficient for various applications. People tend to share their opinions regarding specific product or service aspects. In the case of customer reviews, it has been noted that customers consistently comment on various aspects simultaneously. Aspect-based sentiment analysis (ABSA) adopts a more comprehensive approach than traditional SA levels, provides more meaningful information about an author's perspective on various product or service elements, and covers detailed opinions (sentiment polarity toward various attributes of entities or aspects mentioned in the text) [1]. ABSA was originally launched in SemEval-2014 [2] and contributed datasets containing annotated restaurant and laptop reviews. The

ABSA tasks in SemEval-2014 did not include full reviews until SemEval-2015 [3]. The dataset of SemEval-2016 did not differ from that of 2015 [4]. ABSA may be divided into four main subtasks. Firstly, aspect term extraction (ATE), also known as aspect identification or opinion target extraction, is a core subtask of ABSA that identifies many aspects mentioned in a given sentence. ABSA's primary responsibility is to extract aspects from the review text. For example, in this review about a restaurant, "the pizza is delicious but expensive", we should extract "pizza" as an aspect term. The aspect can be explicit or implicit.

The explicit aspect presents the aspect terms explicitly in the text (noun or phrase noun). In contrast, in the implicit aspect, terms do not appear for the features in the given text. Secondly, aspect polarity classification (APC) involves the classification of opinions regarding various aspects into categories such as "positive", "negative" and "neutral" after the extraction of aspect terms. Thirdly, aspect category detection (ACD) groups synonymous and aspect phrases into aspect categories, where each aspect category represents a particular aspect. In the example sentence, "I must remark that they have one of the fastest delivery times in the city.", the aspect term is "delivery time". For instance, we can group aspect terms with similar meanings into categories where each category represents a particular aspect. For instance, we can group "delivery time", "waiter", and "staff" under the service. Fourthly, aspect category polarity classification (ACPC) performs the task of assigning sentiment polarities ("positive", "negative", and "neutral") to opinions expressed about various aspect categories. This is performed after identifying the aspect category associated with each aspect term.

Multi-task learning is a machine learning method where multiple interconnected tasks are learned simultaneously. This can be beneficial for text classification tasks because it can help a model to learn more generalizable features that are relevant to all of the tasks [5]. Most previous studies addressed the Arabic ATE and ACD tasks separately or sequentially, with separate models developed for each task. In addition, the current Arabic ABSA (AABSA) approaches utilize traditional machine learning (ML) techniques to address AABSA tasks, which require significant time and effort for feature extraction. In addition, the majority of the existing literature on AABSA mainly relies on traditional deep learning approaches, which demand an extensive amount of data to effectively train models.

Additionally, it is important to mention that most of the proposed deep learning-based approaches have relied on traditional word embedding models to generate word vector representations. These models offer static embedding vectors for individual words, regardless of their context. Using transfer learning has the potential to significantly reduce the amount of labeled data and computing resources needed to train a model for downstream tasks. However, there has been a limitation in research utilizing pre-trained language models [6] for AABSA.

Therefore, the goal of this study was to overcome the shortcomings mentioned above by developing a multi-task model called MTL-AraBERT. This model combines AraBERT, BiGRU or BiLSTM, and FNN to extract aspect terms and recognize aspect categories simultaneously.

To the best of our current understanding, this study represents the initial attempt to employ a multi-task learning framework to simultaneously identify both the aspect term and aspect category at the same time in Arabic reviews. In addition, we developed sentence pair classification based on AraBERT with a deep learning model for aspect sentiment polarity classification and aspect category sentiment polarity classification.

The key contributions of this paper are as follows:

1. We propose a multi-task learning model (MTL-AraBERT) that can stimulate ATE and ACD. It integrates the Arabic language model (AraBERT) for contextualized text representation and BiLSTM or BiGRU as a deep layer for extracting more semantics. To the best of our knowledge, this is the first study to develop a multi-task learning model (MTL) for Arabic ATE and ACD.
2. We developed a model based on sentence pair classification (SPC-BERT) with BiLSTM/BiGRU for APC and ACPC.

3. We validated the effectiveness of our suggested models using the publicly available benchmark dataset.

This paper adopts the following structure: Section 2 presents and discusses related work. Section 3 introduces the proposed models. The fourth section details the experiment setup. Section 5 analyzes the results and assesses the models' performance. Finally, in Section 6, the conclusion and future work are discussed.

2. Related Work

We categorized ABSA approaches into early methods, traditional machine and deep learning methods. Modern deep learning utilizes pre-trained language models based on transformers. Rule-based and lexicon-based approaches are handcrafted methods where these approaches depend on handmade rules, lexicons, and linguistic patterns to determine aspect categories or their sentiment polarities [7,8]. Various machine learning techniques have been utilized for ABSA, such as support vector machines (SVMs), naïve Bayes, and random forest classifiers. In addition, hand-crafted features such as n-grams, bag of words (BoW), and term frequency–inverse document frequency (TF-IDF) have been employed. Furthermore, several hybrid approaches combine two or more of these techniques [9].

Deep learning models have enhanced performance by avoiding heavy feature extraction work and automatically learning semantic and syntactic features. Ruder [10] used a simple CNN model to classify aspect categories; it used pre-trained word embedding (Glove) for text representation as input. Then, it used a simple CNN layer with different filters. Xue [11] proposed a CNN network with a gate mechanism to identify aspect categories and sentiment features. Kumar and Ibrahim [12] proposed utilizing word embedding with long-term memory (LSTM) and GRU networks for ACD. In [13], the authors proposed a deep model, namely, the conventional attention-based BiLSTM, considering the next sequence and sentence for ACD in the English reviews dataset for the restaurant industry. In [14], the authors used an attention mechanism to detect the aspect categories from reviews. In [15], the authors proposed a hybrid model by integrating CNN and staked Bi-LSTM with multiplicative attention mechanisms for ACD and ACSC. The attention mechanism has been utilized in ABSA [16] due to its capability to effectively capture the significant components related to a given aspect. In [17], the authors proposed integrating the attention mechanism with LSTM for identifying the important part of a sentence to address the APC task. Another study [18] proposed a BiLSTM model with self-attention for aspect polarity extraction.

Most recently, transformer-based approaches based on contextual pre-trained and transfer learning have been applied to enhance the performance of ABSA. For example, Zhang et al. [19] proposed combining the BERT model with multiple attention layers to improve the performance of ACD. They evaluated the model on Sem-eval 2014. Liao et al. [20] used RoBERTa (robustly optimized BERT pre-training approach) for contextual feature representation and combined it with 1D-CNN and cross-attention for aspect category classification. In [21], the authors performed different pre-trained language models (monolingual and multilingual) on the Vietnamese language.

ATE and ACD are related tasks. As a result, some researchers have suggested methods that can generate both ATE and ACD simultaneously. One such approach is the MTL model, which was proposed by Wei et al. [22]. The model employed multiple layers of CNN for high-level word representations and knowledge propagation between both tasks. Then, a fully connected layer was applied for information extraction. In [23], the authors proposed an MTL model based on question–answering (Q-A)-style reviews and CRF for ATE. In [24], the authors proposed an MTL deep model for ATE and ACD. They considered ATE a sequence labeling problem and used CNN for ATE, and they considered ACD a supervised classification problem and used BilSTM for ACD.

As the proposed solutions for English ABSA, the proposed solutions for AABSA can be categorized as early solutions, including rule-based, lexicon-based, and traditional ML; current solutions, including traditional deep learning and transformer-based models; or a

hybrid of these techniques. A list and comparison of prior research that addresses AABSA are presented in Table 1. A comparison is made between these studies about the AABSA tasks, proposed models, dataset domains, and whether an MTL was implemented.

Table 1. Summary of existing AABSA methods and ABSA tasks (ATE, APC, ACD, ACPC). Note: (×: Doesn't use MTL, ✓ : Use MTL)

Reference	Evaluated Task	Dataset Domain	Proposed Models	MTL	Limitation
[25]	-ATE	-News	-Part of speech, -n-gram for ATE -K-nearest neighbors CRF, decision trees, and naïve Bayes for APC	×	-Require significant time and effort for feature extraction.
[26]	-ACD -APC	-Hotels	-Naïve Bayes, decision trees, and K-nearest neighbors	×	-Require significant time and effort for feature extraction.
[27]	-ATE -APC	-Airline dataset	-Word embedding for feature capturing -SVM for classification	×	-Generate a static embedding vector for each word without capturing the entire context of a word's use.
[28]	-ACD -ATE -APC	-Hotels	-RNN -SVM	×	-Demand an extensive amount of data to effectively train the deep model.
[29]	-ATE -APC	-Hotels	-Bi-LSTM-CRF for ATE -LSTM for APC	×	-Generate a static embedding vector for each word without capturing the entire context of a word's use. -Demand an extensive amount of data to effectively train the deep model.
[30]	-ATE -APC	-Hotels	-Bi-GRU-CNN-CRF for ATE -Interactive attention and GRU for APC	×	-Generate a static embedding vector for each word, regardless of its context. -Demand an extensive amount of data to effectively train the deep model.
[31]	-ATE -ACD	-Arabic news	-BERT-BiLSTM-CRF	×	-Avoid considering the relationship between ATE and ACD. Instead, address these tasks using separate models.
[32]	-ATE	-Hotels	-Rule-based and ontology	×	-Require feature extraction methods. -Require a labeled dataset.
[33]	-ATE	-Hotels	-BERT-Flair-BiLSTM -CRF -BERT-Flair-BiGRU-CRF	×	-Handle only individual tasks (ATE) without considering related ABSA tasks.
[34]	- APC -ACPC	-Book reviews -Hotels	-Sequence to sequence based on BERT	×	-Avoid considering the relationship between ATE and APC.
[35]	-APC	-Book reviews -Hotels -News	-BERT with liner layer for APC	×	-Avoid considering the relationship between ATE and APC.
[36]	-ACD	-Arabic news	-BERT and temporal conventional network and BiGRU	×	-Avoid considering the relationship between ATE and ACD.
[37]	-ATE -APC	-Hotels -Augmented datasets	-MTL model based on LCF-APTEPC and AraBERTvo2	✓	-Require utilizing the model to consider ATE and ACD tasks.

We developed a multi-task learning model based on AraBERT and SPC-BERT model-with deep learning to overcome some of the limitations mentioned above. To the best of our current understanding, our proposed MTL-AraBERT model is the first to utilize multi-task learning-based BERT and deep learning for Arabic ATE and ACD. Our earlier multi-task model [37] focused on ATE and APC tasks. Additionally, there is no prior research that has employed a combination of SPC-BERT and deep learning methods (BiLSTM/BiGRU) for Arabic APC and ACPC (more details in Section 3).

3. Materials and Methods

The proposed model's primary goal was to identify aspect terms and categorize them for Arabic hotel reviews. For example, take this review: "The view from the hotel is excellent; however, the staff is not courteous". In this case, the MTL model should identify the aspect terms "view" and "staff", as well as their corresponding categories, namely, location and service.

Here, a review sentence was defined by the word sequence ($s = w_1, w_2, \ldots, w_N$), where N is the sentence length and w_i is the i-th word in the sentence. The ATE task involved labeling sequences of words using the beginning–inside–outside (BIO) tagging scheme. We assigned one of these labels to each word: B-ASP, I-ASP, or O. B-ASP, indicating the first word of an aspect term; I-ASP, indicating a word inside an aspect term; or O, indicating a non-aspect word. An aspect term could be a single word or a phrase consisting of multiple words.

A review sentence usually covers multiple aspects of the subject being reviewed. Therefore, the ACD task involved classifying the sentence into one or more categories based on the different aspects it covered. This was essentially a multi-label classification problem. In our work, we identified seven categories: Location, Rooms, Services, Food–Drinks, Hotel, Facilities, and Room Amenities. We needed to identify these categories from the sentence as well as from a series of words ($s = w_1, w_2, \ldots, w_N$).

The input sequence was tokenized, and each token was labeled to indicate whether it belonged to the aspect term or its category. Aspect term polarity classification (APC) aimed to classify sentiment polarity for aspect terms (positive, negative, or natural). However, aspect category sentiment polarity classification (ACPC) aimed to classify sentiment polarity for aspect categories.

As an example, Figure 1 shows a review annotated in Arabic: "إطلالة الفندق رائعة والطعام لذيذ لكن الخدمة سيئة" (translates to "the hotel view is great, the food is delicious, but the service is bad"). In this review, the reviewer mentions three aspect terms: "hotel view", "food", and "service". These aspects correspond to three broader categories, also known as aspect categories: "Location", "Food", and "Service". The reviewer expresses positive sentiments toward the view and food, but negative sentiments toward the service.

Review	سيئة Bad	الخدمة service	لكن but	لذيذ delicious	والطعام food	رائعة great	الفندق hotel	إطلالة view
Aspect labels	O	B-ASP	O	O	B-ASP	O	I-ASPl	B-ASP
Aspect category	O	Service	O	Food	O	O	Location	Location
Sentiment polarity	Natural	Negative	Natural	Natural	Positive	Natural	Positive	Positive

Figure 1. An annotated review from Arabic hotel dataset.

In the following subsection, we describe the two proposed models for AABSA tasks. The first model is an MTL model based on AraBERT (MTL-AraBERT) and deep learning methods (BiLSTM/BiGRU) for ATE and ACD. The second model extracts the sentiment polarity for either the aspect term or the aspect category. This is achieved by the implementation of the pair sentence classification model (BERT-SPC) in conjunction with BiLSTM/BiGRU.

Utilizing AraBERT addresses ambiguous sentiment by leveraging its contextual understanding capabilities. This contextual understanding enables AraBERT to take into account the surrounding words and sentences when analyzing sentiment, allowing it to better interpret nuanced or uncertain sentiments in text.

3.1. MTL-AraBERT Model Architecture for ATE and ACD

The proposed MTL model addressed two sub-tasks in Arabic ABSA and simultaneously handled ATE and ACD tasks. Figure 2 illustrates the layers of the MTL-AraBERT architecture.

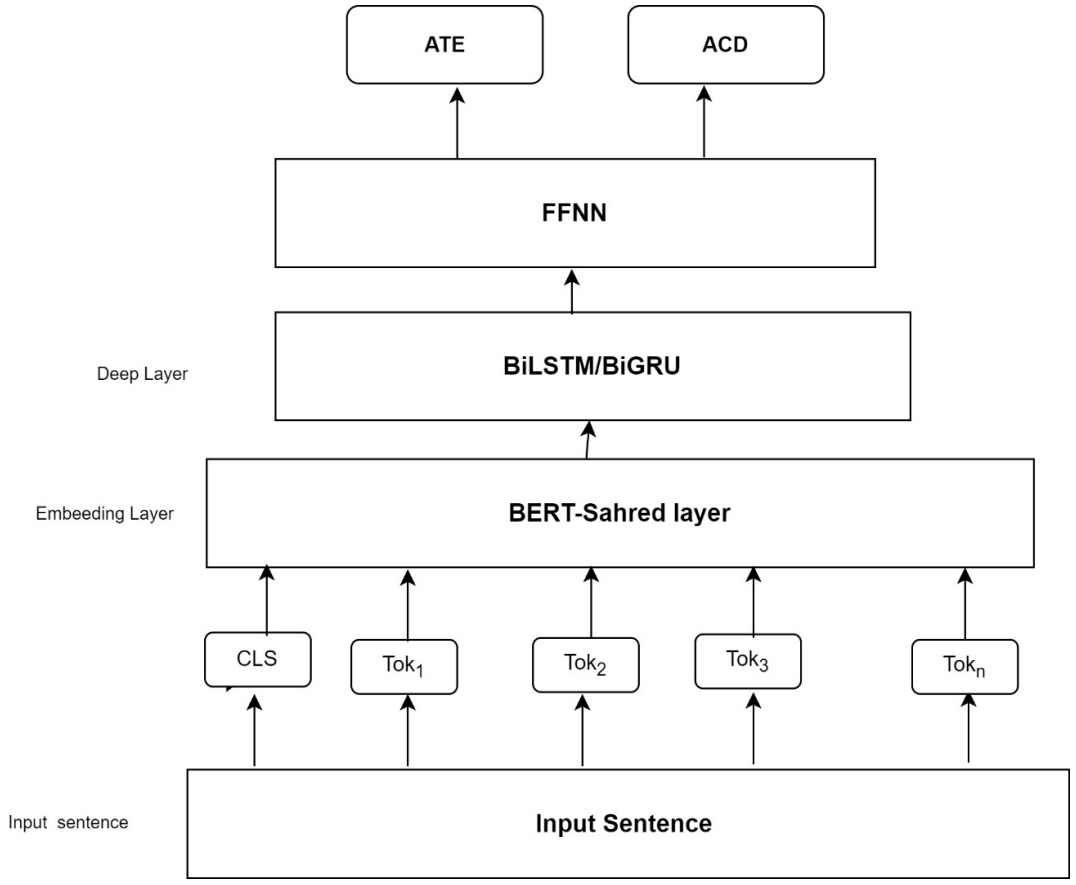

Figure 2. The proposed MTL model architecture combines AraBERT with BiLSTM and BiGRU.

3.1.1. AraBERT Layer

Our proposed approach used AraBERT, a pre-trained Arabic language model, as a shared layer to extract semantic and syntactic features from aspect terms and text in sentence reviews. AraBERT utilized the encoder part of the transformer as a contextual neural encoder to generate contextualized embeddings. This model was trained on a large text corpus (543 MB). The input sentence had a target and category and was given to AraBERT. Then, the sentence was tokenized, and the two special tokens [CLS] and [SEP] were added to the beginning and end of the sentence, respectively. Using WordPress, each input was tokenized. For instance, the review "The hotel is good" was segmented into individual tokens and subsequently processed by the embedding layer. The tokens were mapped into contextualized embedding vectors. The BERT encoder generated a 768-dimension vector for each token in the input. The deep layer, BiLSTM/BiGRU, received each token represented by a 768-dimensional vector. The generic pre-trained BERT model (ARABERT-base) was utilized rather than the large version (AraBERT-large) due to limited computational resources.

The output of BiLSTM/BiGRU was utilized as input for two distinct feed-forward neural network (FFNN) models. These models were employed to perform two tasks: aspect term identification and category detection.

Using AraBERT made the MTL model more general. In addition, the knowledge from ATE and ACD was shareable, which meant that learning about the aspect term from the model was supported by learning about the aspect category task, and vice versa. In addi-

tion, the ATE task focused on the local semantic meaning of each part of a sentence, while the ACD task considered the global features of the entire sentence. Thus, the interaction between these local and global features provided valuable insights for information extraction.

3.1.2. Deep Layer

In this layer, we used two deep learning networks, BiLSTM or BiGRU. We utilized them to capture both forward and backward long dependencies among words in a sequence. These networks could also be stacked on top of each other to build a deep BiLSTM/BiGRU layer. The BiLSTM/BiGRU layer was applied to the last layer of the AraBERT model and returned the scores (probabilities). The last cell of BiLSTM/BiGRU was passed into two different FNNN layers to classify aspect terms and aspect categories.

The loss function calculated a loss using the class probabilities generated by each feed-forward neural network (FFNN) and their respective targets from the dataset. The cross-entropy loss function was used due to its efficacy in handling multi-class classification, complementing the probabilistic outputs of the softmax activation in the model's final layer [38]. However, BCEWithLogitsLoss is a suitable loss function for multi-label classification. The loss functions for ATE and ACD are defined as follows:

$$L_{ATE} = -\sum t_i.\log(p_i) \tag{1}$$

$$L_{ACD} = -w_n[t_i.\log(\sigma(p_i)) + (1 - t_i).\log(\sigma(p_i)) \tag{2}$$

$$L_{ATECD} = L_{ATE} + L_{ACD} \tag{3}$$

L_{ATECD} means the joint loss function, where L_{ATE} means the loss function of ATE; L_{ACD} represents the loss function of ACD.

The MTL-AraBERT model for ATE and ACD works as follows:

1. The dataset consists of sentences, each containing an aspect term and a category. For example, "The pasta is delicious". The aspect term in this sentence is "pasta" and its category is "food".
2. First, a tokenizer is utilized from a pre-trained AraBERT model to generate tokens for each word in the sentence. Then, for word representation, the AraBERT encoder generates an embedding vector for each token in the input sequence vectors, called an embedding vector (with a 768-dimension vector for each token of the sentence).
3. The resultant embedding vectors are then passed through the BiLSTM layer/BiGRU layer, which encodes the contextual information for a specific input word. This step helps the model capture even deeper contextual information for each word.
4. The output from the BiLSTM/BiGRU layer is given to separate neural networks (FFNNs) for two tasks:
    a. Identifying the specific aspects mentioned in the sentence (e.g., "food" in "The food was delicious").
    b. Classifying the aspects into broader categories (e.g., "food" belongs to the "Food" category).

Each FFNN analyzes the information and outputs a probability score for each possible class (aspect term or category). Finally, the most likely class is chosen as the final output.

3.2. BERT-SCP with Deep Learning Methods for APC and ACPC

Here, we propose a model based on auxiliary sentences, where a single sentence is transformed from single sentence classification into a sentence pair classification (SPC) problem [39]. BERT-SPC was utilized based on AraBERT for the aim of sentiment polarity classification, specifically for aspect terms (APC) or pre-determined aspect categories (ACPC), i.e., positive, negative, or neutral. Sentence pair classification (SPC) based on the AraBERT approach was used. The input for the models was paired (sentence, aspect term)

or (sentence, aspect category) for APC and ACPC, respectively. The sentence reviewed the input for AraBERT as follows:

$$\text{Input-APC} = ([\text{CLS}], w1, w2, \ldots wn, [\text{SEP}], a1, a2, \ldots, an, [\text{SEP}])$$

$$\text{Input-ACPC} = ([\text{CLS}], w1, w2, \ldots wn, [\text{SEP}], c1, c2, \ldots, cn, [\text{SEP}])$$

where Input_APC is the input for APC and Input_ACPC is the input for ACPC. [w1, w2, . . ., wn] are words in the sentence, including aspect terms and aspect categories; [a1, a2, . . ., an] are the aspect terms; and [c1, c2, . . ., cn] are the aspect categories relative to the aspect terms.

If a sentence has multiple aspect terms or aspect categories as input pairs, the sentence is repeated for each distinct aspect term or aspect category.

The last hidden layer from the BERT-SPC model, including [CLS], was passed into the deep BiLSTM/BiGRU layer (as explained in Section 3.1.2). Next, the output was fed into a typical FFNN to identify the sentiment polarity for the aspect term or category.

4. Experiment Setting

This section describes the dataset details, the experiment setup to implement the proposed models, and the performance evaluation metrics used to evaluate our research.

4.1. Dataset and Experiment Setting

We conducted experiments using the Arabic hotel reviews dataset from SemEval-2016 task 5 [4]. The dataset has two files, namely, the training data, which consists of 4802 sentences, and the testing data, which consists of 1227 sentences. For the training set, we allocated 90% for training data and 10% for the validation set. Each file contained user reviews and target labels for all ABSA tasks. There were a total of 35 predefined categories. The category for the ACD task comprised an entity type E (such as HOTEL, SERVICE) and an attribute type A (such as PRICE, QUALITY). Each sentence could be associated with several categories. We considered the main seven categories [LOCATION, ROOMS, SERVICE, FOOD_DRINKS, HOTEL, FACILITIES, ROOMS_AMENITIES].

For the MTL model, we made some changes to the original dataset. Every token in the sentence now had two labels. For ATE, we used the BIO annotation strategy (as described in Section 3.1). We also assigned one of the eight aspect categories to each aspect, as mentioned above.

We conducted all tests using the Google Colab Pro+ platform and carried out all implementations using Python 3.12.3 libraries (https://www.python.org/downloads/source/). The contextual embedding layer utilized in our study was the BERT-base-arabertv02-twitter version.

For the models' hyperparameter configuration, we utilized the training dataset for training our model and exploited the validation dataset to choose the hyperparameters. We set a maximum input sequence length of 140, truncating sequences greater than this length and padding sequences less than this length to maintain the same length. We chose the Adam optimizer for improved learning due to its robust performance across a variety of NLP tasks and its efficiency in handling sparse gradients. We selected a learning rate of 2×10^{-5}, aligning with established best practices for fine-tuning BERT models, and a warm-up schedule, gradually escalating the learning rate to our target and thus fostering stable convergence. We capped the training at 15 epochs. To prevent overfitting, we implemented an early stopping mechanism (after every epoch) that monitored validation loss to halt training when performance plateaued. We set a batch size of 16. We set a dropout rate of 0.3 to prevent overfitting [40]. The hyperparameters are shown in Table 2.

Table 2. Experiment hyperparameter settings.

Parameter	Values
Learning rate	2×10^{-5}
Mini-batch size	16
Max. number of epochs	15
LSTM hidden size	256
Hidden size	768
Max. sequence length	140

We set the hyperparameters through a trial-and-error process and tuned them using validation sets—a manual inspection of multiple results within a range identified through existing best practices. We selected the best model hyperparameters based on the best F1 score on the validation set. We experimented with unseen test sets afterward to make sure that we were not overfitting on validation sets.

4.2. Evaluation Metrics

We used the *F*1-score metric to evaluate the performance of ATE and ACD. The *F*1-score was calculated as the harmonic mean of precision (*P*) and recall (*R*). *P* referred to the ratio of accurately predicted aspect terms or aspect categories out of all the predicted ones. On the other hand, *R* represented the ratio of accurately predicted aspect terms or aspect categories to the total number of aspect terms or aspect categories in the original dataset. The *F*1-score was calculated using these two variables.

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \tag{4}$$

We also used accuracy as the evaluation metric for the aspect sentiment polarity classification at the term and category levels. Accuracy was obtained by dividing the number of successfully identified reviews by the total number of reviews.

$$Accuracy = \frac{number\ of\ correct\ samples}{total\ number\ of\ samples} \tag{5}$$

5. Results and Discussion

To evaluate the effectiveness of the proposed models, several experiments were conducted to determine the performance of each model. The performance of each model was compared with relevant works in the literature. The following subsections present and analyze the performance evaluation of the proposed multi-task model (MTL-AraBERT) for ATE and ACD and the performance evaluation of the BERT-SPC model with deep learning for APC and ACPC.

5.1. Performance of AraBERT-Based Multi-Task Learning Model and Deep Learning for ATE and ACD

This subsection presents the performance of MTL-AraBERT on ATE and ACD. A series of experiments were performed to evaluate the effectiveness of our proposed models, namely, MTL-AraBERT, MTL-ArabBERT-BiLSTM, and MTL-AraBERT-BiGRU, in addressing the Arabic ATE and ACD tasks using publicly available benchmark datasets. The models' performance was evaluated using precision, recall, and F1 measures. (A higher F1-score indicated a better model.)

As seen in Table 3, MTL-AraBERT achieved the best F1-score for ATE and ACD tasks among all three proposed models. Adding the BiLSTM and BiGRU layers improved performance compared to previous models but less so than for MTL-AraBERT, where MTL-AraBERT-BiGRU achieved a comparable result with the MTL-AraBERT model on the ATE task. MTL-AraBERT-BiLSTM achieved the worst performance among the three models for the ATE task. However, MTL-AraBERT-BiLSTM achieved the lowest result

among the three models for the ACD task. Overall, the combination of MTL and AraBERT produced a superior result.

Table 3. Performance of MTL-AraBERT.

Model	ATE			ACD		
	Precision	Recall	F1	Precision	Precision	F1
MTL-AraBERT	80.96	79.70	**80.32**	66.96	69.59	**68.21**
MTL-AraBERT-BiLSTM	80.13	77.95	79.01	65.14	65.68	65.33
MTL-AraBERT-BiGRU	79.46	78.27	78.77	66.12	64.58	64.75

The experiment results illustrated in Table 3 show that the MTL-AraBERT model achieved the best results. Thus, to assess the effectiveness of our MTL-AraBERT model, we conducted a comparison between the proposed models and recent approaches that used traditional DNNs: recurrent neural networks (RNNs) [29], BiLSTM-CRF with static pre-trained word embedding, an attention-based neural model approach with pre-trained word embedding [41], and AraBERT [42].

The comparison results of MTL-AraBERT vs. previous deep-based and transformer-based approaches that handle only a single task in one model (on the same benchmark dataset) are presented in Table 4; unreported experiment results from previous works are denoted by "-". As shown in Table 4, the MTL-AraBERT model outperformed all the previous models on a single task. The MTL-AraBERT demonstrated notable performance improvements, with an 18% increase in the F1-score for ATE and a 1.2% increase for ACD. These results provide evidence for the efficacy of the MTL-AraBERT model in enhancing overall performance. By training on multiple related tasks, the multi-task model could learn underlying features and patterns that were beneficial for both tasks. This shared knowledge representation improved the model's ability to identify aspect terms (ATE) and classify their categories (ACD), resulting in improved performance of both ATE and ACD tasks compared to single-task models.

Table 4. Performance of MTL-AraBERT model vs. previous deep-based models.

Model	ATE	ACD
	F1	F1
Bi-LSTM-CRF utilizing wor2vec word embedding [29]	66.32	-
Bi-LSTM-CRF utilizing fast text [29]	69.9	-
Attention-based [41]	72.8	-
BERT-Flair-BISTM-Flair [33]	79.9	-
INSIGHT [10]	-	52.11
C-IndyLSTM [43]	-	58.05
AraBERT [42]	-	67.3
Our proposed MTL-AraBERT	**80.32**	**68.21**

In addition to enhancing the semantic representation of words and relationships in text, AraBERT handles polysemy and misspellings. AraBERT uses a WordPiece tokenizer that splits unknown words into a set of sub-words, which enhances AraBERT's ability to handle OOV issues more effectively.

5.2. Performance of BERT-SPC Model with Deep Learning for APC and ACPC

Here, we present the results of the BERT-SPC-BiLSTM/BiGRU models on APC and ACPC tasks; several experiments were conducted on the Arabic benchmark dataset. The proposed models' performance was evaluated based on F1-score and accuracy.

The results of the comparison between the BERT-SPC model and previous models are summarized in Tables 5 and 6; unreported experiment results from previous works are denoted by "-". As shown in Tables 5 and 6, BERT-SPC-BiGRU with one layer achieved the best performance, with an 89.36% F1-score and 89.67% accuracy for APC. The results indicate that using two layers of BiLSTM or BiGRU significantly improves the model's performance for ACPC, achieving an F1-score of 88.96% and accuracy of 89.36%. Combining BERT-SPC with BiGRU improves feature extraction by considering the appropriate semantic meaning.

Table 5. Performance of the proposed SPC-BERT-BILSTM/BiGRU model on APC vs. existing methods.

Model	Precision	Recall	F1	Accuracy
ASP using BERT	-	-	-	89.51
GRU for APC [35]	-	-	-	83.98
Sequence to sequence transformer [34]	-	-	-	84.65
GRU for APC	-	-	-	83.98
BERT-SPC -BiLSTM/1 layer	88.07	87.58	87.79	88.17
BERT-SPC-BiGRU/1 layer	89.55	89.20	89.36	**89.67**
BERT-SPC-BiLSTM/2 layer	88.07	86.69	87.19	87.71
BERT-SPC-BiGRU/2 layer	89.22	88.20	88.60	89.02

Table 6. Performance of SPC-BERT-BILSTM/BiGRU model on ACPC vs. existing methods.

Model	Precision	Recall	F1	Accuracy
Sequence to sequence transformer [34]	-	-	-	76.48
BERT-SPC-BiLSTM/1 layer	88.07	86.69	87.19	87.71
BERT-SPC-BiGRU/1 layer	89.22	88.20	88.60	89.02
BERT-SPC-BiLSTM/2 layer	87.39	88.09	87.61	87.79
BERT-SPC-BiGRU/2 layer	89.56	88.57	88.96	**89.36**

When we compared our work to DNN-based and transformer-based approaches for APC and ACPC specifically, we compared the proposed model (BERT-SPC-BiGRU) with ASP using BERT [35], GRU for APC [30], and Seq-Seq transformer-based [34] for APC and ACPC. The results show that the proposed model achieved the best performance in terms of F1-score (89.36) and accuracy (89.76). This demonstrated the effectiveness of using BERT-SPC for text representation and adding BiGRU for capturing the long dependencies in text. The proposed model is superior to previous state-of-the-art models because it incorporates AraBERT for word representation, allowing it to use semantic similarities in word embeddings to capture context and identify sentiments for aspects of a text. Another reason for the improved performance is the utilization of AraBERT with sentence-pair classification to calculate similarities between an aspect and a text; it also captures the relationship between an aspect and a sentiment in a text.

Overfitting is a common problem in machine learning and neural network models. Models can easily memorize the specifics of the training data and fail to perform well on new examples. We primarily used dropout and early stopping techniques to prevent overfitting. Dropout is a critical technology for addressing the problem of overfitting in neural networks. The main idea behind dropout is to randomly eliminate a unit and its connection to the neural network during training. This helps to prevent the units from depending too much on each other. In addition, we incorporated an early stopping rule

in our model that checked the validation dataset after each epoch. This validation set was not used during training. At every epoch, the loss function value was calculated on this validation set. The training process was stopped when the validation loss reached its minimum value.

6. Conclusions

This work developed two models to handle the AABSA of the Arabic hotel reviews dataset. These models utilized and integrated state-of-the-art methods based on transforms, transfer learning, deep learning, and multi-task learning. We proposed an MTL-AraBERT model to perform two tasks: aspect term identification and category detection. We fine-tuned AraBERT for contextualized embedding and included a deep layer (BiLSTM or BiGRU) with an FFNN to achieve the specific tasks of ATE and ACD. The results indicate that the suggested model outperforms previous models in terms of aspect extraction and aspect category performance. The proposed approach has the advantage of being able to simultaneously learn two tasks. This is achieved by sharing information and leveraging the interdependencies between these tasks. The model performs better overall compared to models that handle these tasks separately, demonstrating the benefits of multi-task learning in aspect-based opinion mining. Our study also proposes models that combine BERT-SPC with deep models (BiGRU and BiLSTM) for polarity classification on aspects of sentiment polarity and category. The experiment results show that using BERT-SPC-BiGRU is more effective for handling these classification tasks. This highlights the potential for further advancements in this area.

For future work, we plan to implement MTL across various domains in Arabic. Additionally, to enhance the handling of language variations, we plan to increase the size and diversity of training data by utilizing data augmentation techniques. Another option is to pre-train a new Arabic model from scratch on a larger pre-training dataset. Moreover, we intend to design triple and quadratic models for performing more ABSA subtasks at the same time, such as ATE, ACD, APC, and ACPC.

Author Contributions: Conceptualization, A.F., M.S., R.S. and O.A.; methodology, A.F., M.S., R.S. and O.A., software, A.F.; validation, A.F., M.S., R.S. and O.A.; formal analysis, A.F., M.S., R.S. and O.A.; investigation, A.F., M.S., R.S. and O.A.; resources, A.F., M.S., R.S. and O.A.; data curation, A.F.; writing—original draft preparation, A.F.; writing—review and editing, M.S. and O.A.; visualization, A.F., M.S., R.S. and O.A.; supervision, M.S. and O.A.; project administration, M.S. and O.A.; funding acquisition, A.F., M.S., R.S. and O.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data used are publicly available. We used the Arabic benchmark dataset from the SemEval-2016 Task 5 Arabic hotel reviews dataset, which is available at https://github.com/msmadi/HAAD (accessed on 1 January 2024).

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

References

1. Banjar, A.; Ahmed, Z.; Daud, A.; Abbasi, R.A.; Dawood, H. Aspect-Based Sentiment Analysis for Polarity Estimation of Customer Reviews on Twitter. *Comput. Mater. Contin.* **2021**, *67*, 2203–2225. [CrossRef]
2. 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 August 2014; Association for Computational Linguistics: Dublin, Ireland, 2014; pp. 27–35.
3. Pontiki, M.; Galanis, D.; Papageorgiou, H.; Manandhar, S.; Androutsopoulos, I. Semeval-2015 Task 12: Aspect Based Sentiment Analysis. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, CO, USA, 4 June 2015; pp. 486–495.
4. Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S.; Al-smadi, M.; Al-ayyoub, M.; Qin, B.; Clercq, O.D.; Pontiki, M.; et al. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, CA, USA, 16 June 2016.
5. Chen, S.; Zhang, Y.; Yang, Q. Multi-Task Learning in Natural Language Processing: An Overview. *arXiv* **2021**, arXiv:2109.09138.

6. Kenton, J.D.M.-W.C.; Toutanova, L.K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2 June 2019; pp. 4171–4186.
7. Schouten, K.; Frasincar, F.; De Jong, F. Commit-P1wp3: A Co-Occurrence Based Approach to Aspect-Level Sentiment Analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, 23 August 2014; pp. 203–207.
8. Kumar, A.; Saini, M.; Sharan, A. Aspect Category Detection Using Statistical and Semantic Association. *Comput. Intell.* **2020**, *36*, 1161–1182. [CrossRef]
9. Kiritchenko, S.; Zhu, X.; Cherry, C.; Mohammad, S. NRC-Canada-2014: Detecting Aspects and Sentiment in Customer Reviews. In Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), Dublin, Ireland, 23 August 2014; pp. 437–442.
10. Ruder, S.; Ghaffari, P.; Breslin, J.G. INSIGHT-1 at SemEval-2016 Task 5: Deep Learning for Multilingual Aspect-Based Sentiment Analysis. In Proceedings of the 10th International Workshop on Semantic Evaluation, San Diego, CA, USA, 16 June 2016; pp. 330–336. [CrossRef]
11. Xue, W.; Li, T. Aspect Based Sentiment Analysis with Gated Convolutional Networks. In Proceedings of the ACL 2018—56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15 July 2018; Volume 1, pp. 2514–2523. [CrossRef]
12. Kumar, A.; Veerubhotla, A.S.; Narapareddy, V.T.; Aruru, V.; Neti, L.B.M.; Malapati, A. Aspect Term Extraction for Opinion Mining Using a Hierarchical Self-Attention Network. *Neurocomputing* **2021**, *465*, 195–204. [CrossRef]
13. Khan, M.U.; Javed, A.R.; Ihsan, M.; Tariq, U. A Novel Category Detection of Social Media Reviews in the Restaurant Industry. *Multimed. Syst.* **2023**, *23*, 1825–1838. [CrossRef]
14. Movahedi, S.; Ghadery, E.; Faili, H.; Shakery, A. Aspect Category Detection via Topic-Attention Network. *arXiv* **2019**, arXiv:1901.01183.
15. Trueman, T.E.; Cambria, E. A Convolutional Stacked Bidirectional LSTM with a Multiplicative Attention Mechanism for Aspect Category and Sentiment Detection. *Cogn. Comput.* **2021**, *13*, 1423–1432.
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5999–6009.
17. Wang, Y.; Huang, M.; Zhu, X.; Zhao, L. Attention-Based LSTM for Aspect-Level Sentiment Classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1 November 2016; pp. 606–615.
18. Xie, J.; Chen, B.; Gu, X.; Liang, F.; Xu, X. Self-Attention-Based BiLSTM Model for Short Text Fine-Grained Sentiment Classification. *IEEE Access* **2019**, *7*, 180558–180570. [CrossRef]
19. Zhang, X.; Song, X.; Feng, A.; Gao, Z. Multi-Self-Attention for Aspect Category Detection and Biomedical Multilabel Text Classification with Bert. *Math. Probl. Eng.* **2021**, *2021*, 1–6. [CrossRef]
20. Liao, W.; Zeng, B.; Yin, X.; Wei, P. An Improved Aspect-Category Sentiment Analysis Model for Text Sentiment Analysis Based on RoBERTa. *Appl. Intell.* **2021**, *51*, 3522–3533. [CrossRef]
21. Van Thin, D.; Hao, D.N.; Hoang, V.X.; Nguyen, N.L.-T. Investigating Monolingual and Multilingual BERT Models for Vietnamese Aspect Category Detection. In Proceedings of the 2022 RIVF International Conference on Computing and Communication Technologies (RIVF), Ho Chi Minh, Ho Chi Minh, Vietnam, 20 December 2022; IEEE: Ho Chi Minh, Vietnam, 2022; pp. 130–135.
22. Wei, Y.; Zhang, H.; Fang, J.; Wen, J.; Ma, J.; Zhang, G. Joint Aspect Terms Extraction and Aspect Categories Detection via Multi-Task Learning. *Expert Syst. Appl.* **2021**, *174*, 114688. [CrossRef]
23. Wu, H.; Cheng, S.; Wang, Z.; Zhang, S.; Yuan, F. Multi-Task Learning Based on Question–Answering Style Reviews for Aspect Category Classification and Aspect Term Extraction on GPU Clusters. *Clust. Comput.* **2020**, *23*, 1973–1986. [CrossRef]
24. Xue, W.; Zhou, W.; Li, T.; Wang, Q. MTNA: A Neural Multi-Task Model for Aspect Category Classification and Aspect Term Extraction on Restaurant Reviews. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Taipei, Taiwan, 27 November 2017; pp. 151–156.
25. Al-Smadi, M.; Al-Ayyoub, M.; Al-Sarhan, H.; Jararwell, Y. An Aspect-Based Sentiment Analysis Approach to Evaluating Arabic News Affect on Readers. *J. Univers. Comput. Sci.* **2016**, *22*, 630–649.
26. Al-Smadi, M.; Al-Ayyoub, M.; Jararweh, Y.; Qawasmeh, O. Enhancing Aspect-Based Sentiment Analysis of Arabic Hotels' Reviews Using Morphological, Syntactic and Semantic Features. *Inf. Process. Manag.* **2019**, *56*, 308–319. [CrossRef]
27. Ashi, M.M.; Siddiqui, M.A.; Nadeem, F. Pre-Trained Word Embeddings for Arabic Aspect-Based Sentiment Analysis of Airline Tweets. In Proceedings of the International Conference on Advanced Intelligent Systems and Informatics, Cairo, Egypt, 1 September 2018; Springer: Cairo, Egypt, 2018; pp. 241–251.
28. Al-Smadi, M.; Qawasmeh, O.; Al-Ayyoub, M.; Jararweh, Y.; Gupta, B. Deep Recurrent Neural Network vs. Support Vector Machine for Aspect-Based Sentiment Analysis of Arabic Hotels' Reviews. *J. Comput. Sci.* **2018**, *27*, 386–393. [CrossRef]
29. 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]
30. Abdelgwad, M.M.; Soliman, T.H.; Taloba, A.I.; Farghaly, M.F. Arabic Aspect Based Sentiment Analysis Using Bidirectional GRU Based Models. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 6652–6662. [CrossRef]
31. Bensoltane, R.; Zaki, T. Towards Arabic Aspect-Based Sentiment Analysis: A Transfer Learning-Based Approach. *Soc. Netw. Anal. Min.* **2022**, *12*, 7. [CrossRef]

32. Behdenna, S.; Fatiha, B.; Belalem, G. Ontology-Based Approach to Enhance Explicit Aspect Extraction in Standard Arabic Reviews. *Int. J. Comput. Digit. Syst.* **2022**, *11*, 277–287. [CrossRef]
33. 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]
34. Chennafi, M.E.; Bedlaoui, H.; Dahou, A.; Al-Ganess, M.A.A. Arabic Aspect-Based Sentiment Classification Using Seq2Seq Dialect Normalization and Transformers. *Knowledge* **2022**, *2*, 388–401. [CrossRef]
35. Abdelgwad, M.M.; Soliman, T.H.A.; Taloba, A.I. Arabic Aspect Sentiment Polarity Classification Using BERT. *J. Big Data* **2022**, *9*, 115. [CrossRef]
36. Bensoltane, R.; Zaki, T. Combining BERT with TCN-BiGRU for Enhancing Arabic Aspect Category Detection. *J. Intell. Fuzzy Syst.* **2023**, *44*, 4123–4136. [CrossRef]
37. Fadel, A.S.; Abulnaja, O.A.; Saleh, M.E. Multi-Task Learning Model with Data Augmentation for Arabic Aspect-Based Sentiment Analysis. *Comput. Mater. Contin.* **2023**, *75*, 4419.
38. Zhou, Y.; Wang, X.; Zhang, M.; Zhu, J.; Zheng, R.; Wu, Q. MPCE: A Maximum Probability Based Cross Entropy Loss Function for Neural Network Classification. *IEEE Access* **2019**, *7*, 146331–146341. [CrossRef]
39. Sun, C.; Huang, L.; Qiu, X. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2 June 2019; Volume 1, pp. 380–385.
40. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
41. Al-Dabet, S.; Tedmori, S.; Al-Smadi, M. Extracting Opinion Targets Using Attention-Based Neural Model. *SN Comput. Sci.* **2020**, *1*, 242. [CrossRef]
42. Ameur, A.; Hamdi, S.; Yahia, S. Ben Multi-Label Learning for Aspect Category Detection of Arabic Hotel Reviews Using AraBERT. In Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023), Lisbon, Portugal, 22 February 2023; pp. 241–250.
43. 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]