knowledge

MDPI

Article

Arabic Aspect-Based Sentiment Classification Using Seq2Seq Dialect Normalization and Transformers

Mohammed ElAmine Chennafi [1,†], Hanane Bedlaoui [1,†], Abdelghani Dahou [1,2,*] and Mohammed A. A. Al-qaness [3,4]

[1] Department of Mathematics and Computer Science, Faculty of Science and Technology, University of Ahmed DRAIA, Adrar 01000, Algeria
[2] Sustainable Development and Computer Science Laboratory Laboratory, Faculty of Science and Technology, University of Ahmed DRAIA, Adrar 01000, Algeria
[3] State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
[4] Faculty of Engineering, Sana'a University, Sana'a 12544, Yemen
* Correspondence: dahou.abdghani@univ-adrar.edu.dz
† These authors contributed equally to this work.

Abstract: Sentiment analysis is one of the most important fields of natural language processing due to its wide range of applications and the benefits associated with using it. It is defined as identifying the sentiment polarity of natural language text. Researchers have recently focused their attention on Arabic SA due to the massive amounts of user-generated content on social media and e-commerce websites in the Arabic world. Most of the research in this fieldwork is on the sentence and document levels. This study tackles the aspect-level sentiment analysis for the Arabic language, which is a less studied version of SA. Because Arabic NLP is challenging and there are few available Arabic resources and many Arabic dialects, limited studies have attempted to detect aspect-based sentiment analyses on Arabic texts. Specifically, this study considers two ABSA tasks: aspect term polarity and aspect category polarity, using the text normalization of the Arabic dialect after making the classification task. We present a Seq2Seq model for dialect normalization that can serve as a pre-processing step for the ABSA classification task by reducing the number of OOV words. Thus, the model's accuracy increased. The results of the conducted experiments show that our models outperformed the existing models in the literature on both tasks and datasets.

Keywords: sentiment analysis; aspect-based sentiment analysis; normalization; Seq2seq; aspect term polarity; aspect category polarity

Citation: Chennafi, M.E.; Bedlaoui, H.; Dahou, A.; Al-qaness, M.A.A. Arabic Aspect-Based Sentiment Classification Using Seq2Seq Dialect Normalization and Transformers. *Knowledge* **2022**, *2*, 388–401. https://doi.org/10.3390/knowledge2030022

Academic Editor: Gautam Srivastava

Received: 28 June 2022
Accepted: 28 July 2022
Published: 4 August 2022

1. Introduction

In recent years, the world has seen significant developments in the web sector and there has been growing interest in social media. Moreover, users produce and generate a considerable amount of data each day. This user-generated content includes essential information about opinions on multiple topics. Therefore, there is a growing need to understand human opinions and feelings to make the process of decision-making on products, places, and services, easy [1]. Sentiment analysis (SA), also known as opinion mining, is a subfield of natural language processing (NLP) that identifies the sentiment of a given text automatically as it is positive, negative, or neutral [2,3].

One of the SA classification levels is aspect-based. Aspect-based sentiment analysis (ABSA) is a more fine-grained and complex task than SA; it is concerned with determining the aspect terms presented in a document, as well as the sentiment expressed against each term [4].

For instance, Figure 1 illustrates an example restaurant review with two different target aspects and their related sentiment expressions. As shown, the first target aspect,

LOCATION, expressed by the aspect expression "view of the river" has a positive polarity. In contrast, the second target aspect, "FOOD" represented by the aspect expression "sushi rolls", has a negative polarity.

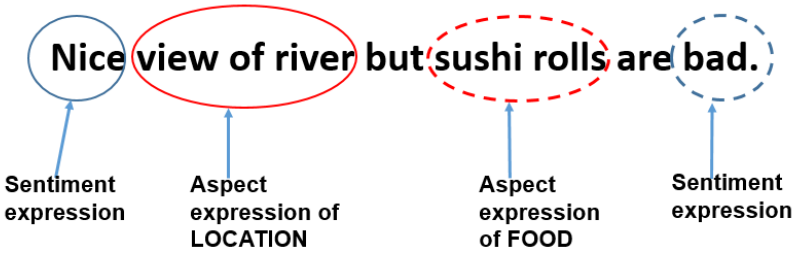

Figure 1. An aspect-based sentiment analysis example. The aspect expressions "view of river" and "sushi rolls" correspond to the sentiment expressions "nice" and "bad".

As mentioned in [5], there are four major tasks of ABSA that can be identified: aspect term extraction (T1), aspect term polarity (T2), aspect category identification (T3), and aspect category polarity (T4). In this work, we concentrate on tasks (T2) and (T4). The difference between aspect term and aspect category is that aspect terms are more precise and should appear in the review sentence, while the aspect category does not appear in the review sentence. Aspect categories are not identified using lateral terms in a sentence but rather inferred by the use of words, adjectives, or the context of the sentence's meaning [6].

Recently, the field of Arabic NLP has become more interesting and challenging [7]. Arabic is the official language of twenty-two (22) countries across Asia and Africa and is spoken by 300 million people. It is one of six official languages of the United Nations [8]. Two Arabic types can be considered: modern standard Arabic (MSA) and dialect (vernacular). MSA originated from classical Arabic and is utilized throughout the Arab world in education, media, literature, official documents, and books. Before the mid-1990s, the only documented versions of Arabic were classical Arabic and MSA. After that, the documentation of different Arabic dialects was pushed due to the widespread use of internet services and mobile applications. In addition, a new version of Arabic known as Arabizi has been designed in which Arabic characters have been replaced with the Roman alphabet [6]. Arabic dialects are used in the daily informal communications between people who live in the same country. Linguists have classified Arabic vernaculars into seven main regional groups, which are Egyptian, Maghrebi, Mesopotamian, Sudanese, Arabian Peninsula, Andalusian, and Levantine [6].

The ambiguity and complexity of Arabic morphology and orthography are the main challenges and problems known by the Arabic SA. The Arabic language has a set of morphemes, such as affixes, prefixes, and suffixes, which express linguistic features, such as a person, instance, and gender. Furthermore, Arabic has a set of 16 morphemes, which adds to the ambiguity caused by diverse meanings of the same word, making Arabic text processing more difficult. As illustrated in Figure 2, the term (فسيرونها) has many morphemes that express an English sentence, "and they will see it" [8,9].

Orthographically, the lack of diacritical markings in the text creates lexical ambiguity, which is a challenging problem for computer systems [8]. For instance, the undiacritized word (درس) may have several meanings, "lesson", "study", and "taught". Moreover, the massive number of Arabic dialects and the unavailability of dialectical Arabic (DA) language resources are well-known problems that lead to the lack of training datasets, making research in this field more complicated. Consequently, these dialectical words are considered out-of-vocabulary (OOV) words in many pre-trained language models.

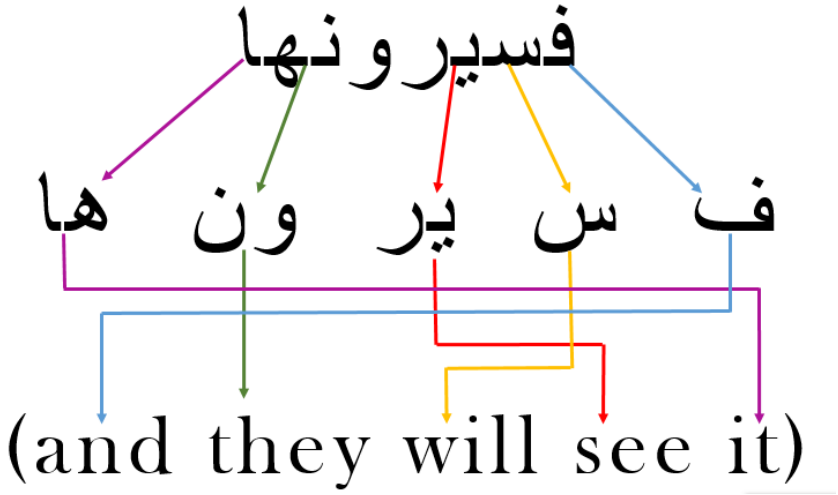

Figure 2. Example of the ambiguity of Arabic morphology.

The role of the normalization phase removes the out-of-vocabulary (OOV) words or the dialectical words and replaces them with suitable forms of modern standard Arabic (MSA) to improve the performance of the aspect-based level.

In this study, we adopted a sequence-to-sequence model for text normalization that aims to transform dialect into MSA. Moreover, we adopted two aspect-based sentiment analysis models that rely on the pre-trained bidirectional encoder representations from transformers (BERT) language model for Arabic to perform aspect-based classifications for both the task aspect category polarity and aspect sentiment classification.

The key contributions of our work are:

- We propose a new solution that improves the results of ABSA by converting the dialectical text into MSA using text normalization.
- We used a pre-trained based model (BERT) with sentence pair input to solve the Arabic ABSA classification task, including MSA and dialect, instead of traditional machine learning (ML) algorithms.
- We adopted a sequence-to-sequence model for normalizing out-of-vocabulary (OOV) words from our dataset.
- We pre-processed and built a training dataset for the normalization model using two well-known public datasets.

The rest of the research is given as follows. Section 2 addresses the recent related work on ABSA. In Section 3, we explain the methodology used in our research. Then, we present the experimental setup in Section 4. Section 5 discusses the results of our experiments. Finally, Section 6 presents the study's conclusions and future work.

2. Related Work

For Arabic dialects, some existing studies addressed different problems. For example, Hamada and Marzouk [10] developed a project called ALMoFseH. The goal of this project was to build a hybrid system that translates most of the dialects on social media into MSA, This system consists of three components: disambiguation of the morphological analysis output using naive Bayesian learning, a rule-based transfer system, and a dictionary look-up system, which offers highly accurate results.

Al-Ibrahim and Duwairi [11] presented a framework for translating the Jordanian dialect into modern standard Arabic (MSA) using the RNN decoder model, and it provided excellent results on a manually generated data set. The word is much better than the results at the sentence level.

Torjmen and Haddar [12] created a translator to translate the Tunisian dialect into MSA. The method consists of developing a set of dictionaries and the construction of inflectional, morphological, and grammatical grammar using finite-state transducers. This

method was implemented and tested using the new technologies provided by the NooJ linguistic platform. The experiments were conducted on two different test sets containing more than 15,000 words, and they presented promising results.

In [13], the authors studied the problem of Arabic text information extraction from social media. They proposed an integrated approach to preprocess Arabic social media data that contained both standard Arabic and dialects. In [14], the authors collected user-generated text for the Moroccan dialect to Moroccan language resource using character neural embedding. They applied normalization and preprocessing techniques to clean the collected data and used deep learning models to analyze the collected data. The authors in [15] studied the impacts of preprocessing on offensive classification for Arabic. They employed several preprocessing and normalization methods for standard Arabic and different dialects. They analyzed the collected data with different classifiers and they concluded that the impacts of preprocessing benefited traditional machine learning methods, but the BERT-based classifiers had no benefit from the preprocessing. In [16], the authors presented a systematic review of the sentiment analysis of existing dialect Arabic studies. They used 60 published papers between 2010 and 2020. They concluded that the support vector machine and the naive Bayes classifier are the most applied techniques to classify Arabic dialects. They also concluded that NLP applications for Arabic dialects need more investigations.

For ABSA, the existing research methodologies can be classified into traditional machine learning techniques and deep learning techniques. ABSA's early efforts were primarily based on deep learning techniques. In this section, we will mention some of the most recent works tackling English and Arabic ABSA, as follows.

2.1. English Aspect-Based Sentiment Analysis

Xue et al. [17] conducted extensive experiments on SemEval datasets provided by [18], showing improved performances of their efficient gated Tanh–ReLU units (GTRU) with the gating mechanisms model compared to other neural models. The proposed model controls the sentiment flow according to the aspect information for aspect category sentiment analysis (ACSA) and aspect-term sentiment analysis (ATSA) tasks. In addition, The aspect and sentiment information are modeled separately by two conventional layers.

Liu et al. [19] introduced a gated alternate neural network (GANN) model, a new type of neural network architecture. Moreover, a gate truncation RNN (GTR) module was used to learn useful aspect-dependent sentiment clue representations. Convolution and pooling algorithms produce more exact representations by using a gating mechanism to control information and extract key local sentiment clue features and obtain the position invariance of features. They experimented with their models using four Chinese and three English datasets and achieved the best results.

Li et al. [20] conducted experiments on two benchmark datasets originating from SemEval [18] and re-prepared in [21] to assess the efficiency of the BERT embedding component on the end-to-end aspect-based sentiment analysis (E2EABSA) task. The experimental results illustrate that their BERT-based model outperforms state-of-the-art works.

Xu et al. [22] improved the accuracy of two tasks, aspect term sentiment analysis (ATSA) and aspect category sentiment analysis (ACSA), by proposing a recurrent neural network model with target embedding (RTE) using the target enhance technique. Experiments were conducted on the SemEval workshop and a Twitter dataset. The proposed RTE model achieved the best results compared to state-of-the-art models.

Trueman et al. [23] evaluated their model based on convolutional stacked bidirectional long short-term memory with a multiplicative attention mechanism using SemEval-2015 and SemEval-2016 datasets on aspect category and sentiment polarity detection tasks.

2.2. Arabic Aspect-Based Sentiment Analysis

The number of studies available in the Arabic ABSA is still restricted due to the language complexity.

Abdelgwad et al. [24] experimented with the modeling capabilities of contextual embeddings from the pre-trained BERT model on the Arabic aspect sentiment polarity classification task with the use of sentence pair inputs. Three different Arabic datasets have been used, including HAAD [5], the Arabic news dataset [25], and Arabic hotel reviews datasets [18]. Their model had the best accuracy of 89.51% on the Arabic hotel reviews dataset.

Abdelgwad et al. [4] addressed aspect opinion target extraction (T2) and aspect polarity detection (T3) tasks by using two recurrent gated unit (GRU)-based models. They used the benchmark Arabic hotel reviews dataset [18]. The proposed methods achieved the best results on both tasks, with an F1 score of 70.67% for Task 2 and an accuracy of 83.98% for Task 3.

Ashi et al. [26] compared two word-embedding models—fastText Arabic Wikipedia and AraVec-Web—for the two-task aspect detection followed by sentiment polarity classification of the detected aspects. They used a corpus of 5K airline service-related tweets in Arabic. A support vector machine (SVM) classifier was used for both classification tasks. They found that the fastText Arabic Wikipedia word embeddings model performed best with an accuracy of 70% for aspect detection and 89% for sentiment polarity classification of the detected aspects.

Al-Dabei et al. [27] presented two deep learning models to cover both ABSA tasks—aspect category identification and aspect sentiment classification. The first one is based on the CNN and the stacked independent LSTM model. The second model comprises multiple layers of stacked bidirectional independent LSTMs, a position-weighting mechanism, and multiple attention mechanisms. They used the Arabic SemEval-2016 dataset for the hotel domains. The two models achieved the best results, with an F1 score of 58.08% for the first model and an accuracy of 87.31% for the second model.

Mohammad et al. [28] developed a deep learning model (Pooled-GRU) based on gated recurrent unit(s) (GRU) and features extracted to handle two ABSA tasks—aspect extraction and aspect polarity classification—by using multilingual universal sentence encoder (MUSE). They employed the Arabic hotel reviews dataset and obtained the best results in both tasks, achieving an F1 score of 93.0% in the first task and 90.86% in the second. Al-Smadi et al. [29] implemented two LSTM-based neural network models. the first was a character-level bidirectional LSTM with a conditional random field classifier (Bi-LSTM-CRF) for aspect opinion target expression (OTA) extraction and the second was an aspect-based LSTM for aspect sentiment polarity classification. The experiments conducted on Arabic hotel reviews enhanced both tasks by 39% for aspect-OTE extraction and 6% for aspect sentiment polarity classification.

Al-Smadi et al. [30] implemented and trained two approaches of RNN and SVM along with the lexical, syntactic, word, morphological, and semantic features, and evaluated using the Arabic hotel reviews dataset. The results showed that the SVM approach was better than RNN in the three ABSA tasks, whereas the deep RNN execution time for training and testing was faster.

Table 1 summarizes several recent related works in English and Arabic aspect-based sentiment analyses.

Table 1. Summary of the related works.

Paper	Task	Model	Dataset
	Aspect-based sentiment analysis in English		
Xue et al. [17]	ACSA and ATSA	CNN and gating mechanism	SemEval2014 datasets
Liu et al. [19]	ABSA	Gated Alternate NN (GANN)	SemEval2014, four Chinese and Tweeter dataset
Li et al. [20]	Aspect term (E2E-ABSA)	BERT	Two review datasets from SemEval
Xu et al. [22]	ATSA and ACSA	RNN and Target embedding	SemEval and a twitter dataset
Trueman et al. [23]	ACSA and SPD	convolutional stacked bi-LSTM and attention mechanism	SemEval-2015 and SemEval-2016
	Aspect-based sentiment analysis in Arabic		
Abdelgwad et al. [24]	ASPC	Pre-trained model BERT	HAAD, Arabic News and Arabic Hotel Reviews datasets
Abdelgwad et al. [4]	ASPC and AOTE	GRU and CNN	Arabic hotel reviews dataset
Ashi et al. [26]	AE and ASPC	Word embedding	Arabic airline-related tweets
Al-Dabet et al. [27]	OTE extraction and ASPC	CNN and LSTM	Arabic SemEval-2016 dataset
Mohammad et al. [28]	AE and ASPC	(GRU)	Arabic hotel reviews
Al-Smadi et al. [29]	Aspect OTE and ASPC	LSTM	Arabic Hotels' reviews
Al-Smadi et al. [30]	ACSA, aspect OTE extraction and ASPC	RNN and SVM	Arabic Hotels' reviews

3. Proposed Model

3.1. Encoder–Decoder Architecture

The encoder–decoder model is divided into two components. First, an encoder that takes an input $\mathcal{X}$ (in this example, a phrase) and generates an intermediate representation $\mathcal{Z}$ (or code) that emphasizes its key features; and second, a decoder that processes that collection of features and generates the needed output $\mathcal{Y}$ (in this case, a normalized phrase). $\mathcal{Z}$ is a matrix with the dimensions $\mathcal{Z} \in \mathcal{M}_{f*l}$, where f is the number of features to encode for each input value [31]. This model's fundamental diagram is shown in Figure 3.

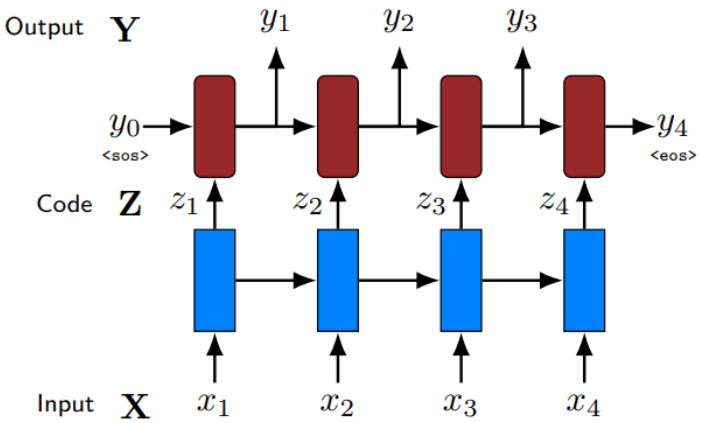

Figure 3. An encoder–decoder architecture.

The following section provides an overview of the overall methodology used in this paper. Our research focuses on the aspect term polarity and aspect category polarity tasks. These two tasks can be formulated as a pair sentence classification task. Figure 4 describes the overall architecture of our model.

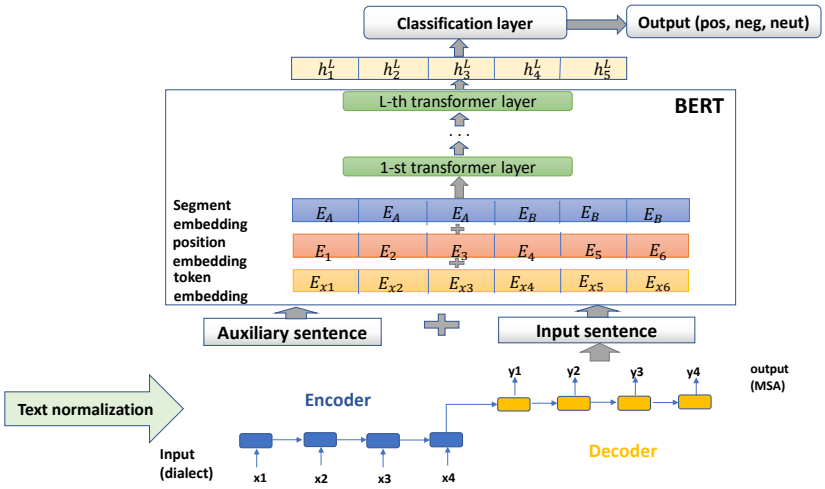

Figure 4. The architecture of the proposed model.

Our normalization model is based on a sequence-to-sequence learning framework introduced by [32]. The model consists of two components—a normalization part in which the dialect text is converted into MSA. The model reads the informal text sequences (dialect text) and transforms them into a continuous-space representation passed on to the decoder to generate the target normalized sequence (MSA). The input and output of this process should have the same length. The input token sequences $x = [x_1, ..., x_T]$ of length T represent the dialect text and the generated output sequence (MSA) $y = [y_1, ..., y_L]$ with length L. The input sequence x is read and transformed by the encoder module into a context-specific hidden state sequence $h = [h_1, ..., h_T]$. The concatenation of the two encoder modules (forward and backward) at time t yields the final hidden state:

$$h_t = [g_f(x_t, h_t - 1); g_b(x_t, h_t + 1)] \quad (1)$$

where g_f and g_b denote the forward and backward encoder units, respectively. Based on the previous word y_{j1} and the decoder state s_{j1}, a hidden state sequence $s_j = g_s(s_{j1}, y_{j1}, c_j)$ is produced by the decoder. The context vector c_j is calculated as a weighted sum of encoder hidden states based on the attention mechanism [33]. Then, the Softmax classifier predicts each target word. Figure 5 presents an example of the source (dialect) and target (MSA) pair of sentences for which the seq2seq model helps in appropriately normalizing the content.

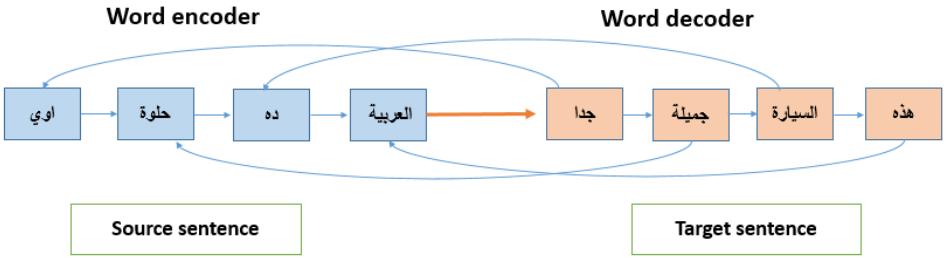

Figure 5. Seq2seq model for dialect normalization.

The second part is the aspect-based sentiment analysis that consists of two pre-trained language models—BERT for aspect term polarity (T2) and the aspect category polarity (T4) accept two sentences as inputs. The first one is the output of the normalization framework (MSA) and the second one is the auxiliary sentence related to the first one.

3.2. Explanation of the T2 Model

The T2 task model involves the aspect and the normalized review sentence mentioning that aspects are the two input sentences. The input sequence is formulated as follows: $x = ([CLS]; a_1, ..., a_m; [SEP]; s_1, ..., s_n; [SEP])$, where $a_1, ..., a_m$ denotes the auxiliary sentence that contains the aspect with m tokens and $s_1, ..., s_n$ represents the review sentence with n tokens that contains that aspect. The polarity distribution is predicted as follows:

$$L_4 = softmax(W_4.h[CLS] + b_4) \quad (2)$$

where $h[CLS]$ is the aspect-aware representation of the whole input and $W_4, b_4 \in R^3$ (3 is the number of polarities). On $[CLS]$, the function is applied along the label dimension: $L_4 \in [0, 1]^3$. The introduction of the post-training step to boost the performance of aspect sentiment classification is required due to the insufficient supervised training data, which limits the performance gain across the tasks.

3.3. Explanation of the T4 Model

To handle the aspect category polarity task T4, a BERT model proposed by Sun et al. [34] was adopted with some modifications to deal with our task. An auxiliary sentence was constructed from the aspect to transform aspect category polarity, which aimed to determine fine-grained sentiment polarity towards a given category associated with a term, into a sentence-pair classification task, The pre-trained BERT model was fine-tuned and evaluated on the HAAD task dataset for aspect category polarity. The final hidden state or (output of the transformer) of the first token was used as input to obtain a fixed-dimensional pooled representation of the input sequence. A classification layer whose parameter matrix was $\mathcal{W} \in \mathcal{R}^{k*H}$ was added, where $\mathcal{K}$ represents the number of categories and the vector $\mathcal{C} \in \mathcal{R}^H$. The softmax function was used to determine the probability of each category $\mathcal{P}$:

$$\mathcal{P} = softmax(\mathcal{C}\mathcal{W}^T) \quad (3)$$

For the T4 model, we used two methods to construct the auxiliary sentence:

- **QA-M method:** It refers to the question-answering task. The auxiliary sentence generated from the category is a question.
  As an example (Table 2), for "رواية رائعة ، انا حسيت وكأني عايش", the category here is "المشاعر", where the generated sentence is "ما رأيك في المشاعر ؟"
- **NLI-M method:** For the natural language inference (NLI) task, the auxiliary sentence contains only the category of the sentence. For the previous example (Table 3), the auxiliary sentence formed is ألمشاعرْ.

For each sentence, the related category is polarized as positive, negative, or neutral, with the other categories being labeled as none. Because the number of categories in the HAAD dataset is 15, we suggested randomly selecting three non-labeled categories instead of all categories to have a balanced dataset.

Example 1. *S1* = كتب يحكي تاريخ الاندلس. رائع بحق

Table 2. An example of the HAAD dataset with the QA-M method for S1.

Sentence	Auxiliary Sentence	Sentiment
S1	ما رأيك في الاسلوب ؟	Positive
S1	ما رأيك في المشاعر ؟	None
S1	ما رأيك في الحبكة ؟	None
S1	ما رأيك في السياق ؟	None

Table 3. An example of the HAAD dataset with the NLI-M method for S1.

Sentence	Auxiliary Sentence	Sentiment
S1	الاسلوب	Positive
S1	الهوامش	None
S1	السياق	None
S1	الخاتمة	None

4. Experimental Setup

This section details the experimental setup used in our research. In our experiments, we investigated the impact of using dialect normalization on the two task models, BERT.

4.1. Dataset Description

We conducted our experiments using four Arabic datasets. PADIC corpus [35] has 6400 sentences for each of the six Maghreb and Middle Eastern dialects (Annaba, Algiers, Tunisian, Morocco, Syrian, and Palestinian), and MSA was used to align each dialect. MADAR corpus [36] is a large parallel corpus that was constructed by translating selected sentences in English and French from the basic traveling expression corpus (BTEC) into the dialects of 25 Arabic cities, in addition to MSA. It contains two corpora: Corpus-26 and Corpus-6. HAAD [5] is considered the first accessible dataset for Arabic ABSA. It contains 1513 Arabic book reviews and 2838 aspect terms. The Arabic hotel reviews dataset [18] was submitted at SemEval-2016 to support ABSA's multilingual task, which covers eight languages and seven domains. The first two datasets are used for the training normalization model, while the latter two are for T2 and T4. A summary of our distribution is presented in the following Tables 4 and 5.

Table 4. Our selected samples from MADAR and PADIC datasets in which the input and output are the same lengths.

Dataset	Total Samples	Selected Samples
PADIC	32,060	16,978
MADAR	100,000	18,994

Table 5. Our sample distribution for HAAD and SemEval-2016 datasets with T2 and T4 models. Pos refers to positive, Neg (negative), Neut (neutral), and conf (conflict).

Dataset	Task Model	Label	Training Set	Valid Set	Test Set
HAAD	T2	Pos	1054	37	285
		Neg	972	52	263
		Neut	118	5	24
	T4	Pos	586	/	137
		Neg	590	/	161
		Neut	16	/	3
		conf	18	/	2
SemEval-2016	T2	Pos	5747	72	1426
		Neg	3119	22	784
		Neut	654	6	162

4.2. Hyper-Parameter Setting

For the normalization framework, we used the Adam optimizer [37] and set the learning rate to 0.1 with a dropout of 0.1 for all experiments. On the other hand, for the BERT models, we used the Adam optimizer and set the learning rate to 3×10^{-5}, with a batch size of 32 and 8 for T2 and T4, respectively. T2 and T4 models were trained for five and four epochs, respectively. The Hugging Face Transformers library, reference [38], was used in all our experiments. Pre-trained language models "Arabic BERT" [39] and "araBERT" [40] were adopted.

4.3. Performance Measures

To validate the strength of the proposed model, multiple models were implemented; the results are compared. The accuracy measure was used to assess the efficacy of the proposed framework, defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

where *TP* (true positive) and *TN* (true negative) are the correctly predicted samples. False positive (*FP*) and false negative (*FN*) are the incorrectly predicted samples by the model. So, accuracy is defined as the number of correct samples of the total number of samples. Better performance is represented by higher accuracy.

5. Results and Discussion

This section presents the experiments carried out as part of our research. Moreover, a comparison between our model and previous works was made.

5.1. Experimental Series 1

In this experimental series, we conducted three different experiments to investigate the effects of using dialect normalization on our task models (BERT). Table 6 summarizes our three experiments:

Table 6. Summary of the conducted experiments. Exp 1, Exp 2, and Exp 3 refer to the first, second, and third experiments, respectively.

Experiment	Normalization Training Dataset	Normalization Training Accuracy	Task Model Training Dataset
Exp 1	Cairo dialect from MADAR	91.1%	HAAD and SemEval with only normalized Egyptian samples
Exp 2	All PADIC and MADAR	95%	HAAD and SemEval with only normalized Egyptian samples
Exp 3	All PADIC and MADAR	95%	All normalized samples from HAAD and SemEval

To obtain Egyptian dialect sentences, we used CAMeL tools [41]. It is a collection of free and open-source tools for processing natural Arabic language developed by the CAMeL Lab. It has several tools for pre-processing, including dialect identification. There are 161 Egyptian dialect sentences in the HAAD dataset and 395 Egyptian dialect sentences in the SemEval dataset

We list the obtained results with and without normalization for the three experiments in Tables 7 and 8 for T2 and T4, respectively.

Table 7. Performance results on T2 using HAAD and SemEval-2016 datasets for the three experiments. Accuracy metric was used.

Dataset	HAAD	SemEval-2016
Without Norm	73.42%	83.76%
Exp 1	**74.85%**	83.81%
Exp 2	74.77%	**84.65%**
Exp 3	66.39%	79.76%

Table 8. Performance results on T4 using NLI-M and QA-M methods for the three experiments. Accuracy metric was used.

Method	NLI-M	QA-M
Without Norm	75.90%	75.08%
Exp 1	**76.89%**	75.49%
Exp 2	76.50%	**76.48%**
Exp 3	75.42%	75.24%

Tables 7 and 8 show the model performances for all experiments. For aspect term polarity, the best-obtained results were with Exp 1 and Exp 2 for HAAD and SemEval-2016 datasets, respectively. For aspect category polarity, the best-achieved results were with Exp 1 and Exp 2 for NLI-M and QA-M, respectively. We noticed that the results improved when we applied normalization only to Egyptian samples. This confirms that Egyptian words are considered out-of-vocabulary words in BERT's vocabulary, and reducing them will give us the best results. Meanwhile, in the third experiment, when we applied normalization to all datasets, we noticed that the performance declined. We can justify this decrease because our normalization model may give us OOV words when we use it on MSA, and then the accuracy decreases.

The results showed good performances of both the aspect term polarity and aspect category polarity models with the first and the second experiments when we applied dialect normalization only on Egyptian sentences in HAAD and SemEval-2016 datasets. Regarding aspect term polarity, the best accuracy of HAAD was with the first experiment (74.85%)

and SemEval-2016 with the second experiment (84.65%). For aspect category polarity, the best accuracy of the NLI-M method was with the first experiment (76.89%) and QA-M with the second experiment (76.48%). This improvement confirms that Egyptian words are considered out-of-vocabulary words in BERT's vocabulary, and reducing them will provide us with the best results. Meanwhile, in the third experiment, when we applied normalization to all datasets, we noticed that the performance declined. We can justify this decrease because our normalization model may give us OOV words when we use it on MSA, and then the accuracy decreases.

5.2. Experimental Series 2

As a final experiment, we compared our best-obtained results of the three experiments and the previous existing works on both aspect term polarity and aspect category polarity. As Table 9 shows, our models achieved the best results on both HAAD and SemEval-2016 datasets with our research tasks T2 and T4.

Table 9. Comparison of our best models with previous works.

Model	HAAD	SemEval-2016
T2: Aspect term polarity		
Abdelgawad et al. [4]	NA	83.98
Abdelgawad et al. [24]	73.23	NA
Al-Smadi et al. [29]	NA	82.6
Ruder et al. [42]	NA	82.7
Our model	**74.85**	**84.65**
T4: Aspect category polarity		
Obaidat et al. [6]	71	NA
Our model	**76.48**	NA

6. Conclusions

Sentiment analysis is one of the most important areas of NLP. Unlike English, a few studies on Arabic sentiment analysis focused on the aspect level. The most important challenge that faces the Arabic ABSA is the lack of dialectal resources and datasets that can be used to train the ABSA model. Furthermore, dialectal Arabic is difficult to process because it breaks all grammatical rules, reducing the ABSA model accuracy. In recent years, pre-trained language models, such as BERT, have shown great effectiveness in sentiment analysis. For Arabic, the small and restricted number of available datasets that cover multiple Arabic dialects reduces the BERT model in-vocabulary words, decreasing the model's performance. This study aimed to increase the effectiveness of the BERT model and reduce the out-of-vocabulary words by translating our dialectical text and transforming it into formal speech. Specifically, the addressed ABSA tasks in this research are aspect term polarity (Task T2) and aspect category polarity (Task T4). Our best results outperformed the previous existing works, where we obtained (in T2) the best results in both datasets, HAAD and SemEval. The accuracy results were 74.85% in the HAAD dataset and 84.65% in the SemEval dataset. In T4, we obtained an excellent result as well (76.84%). Future goals involve generalizing our study on the rest of the aspect-based sentiment analysis tasks and improving the normalization performance by focusing on more dialectical Arabic. In addition, experimenting with models that combine several DL architectures, such as recurrent neural networks and convolution neural networks, may boost the performances on Arabic tasks.

Author Contributions: Conceptualization, A.D.; methodology, M.E.C., H.B. and A.D.; software, M.E.C. and H.B.; validation, A.D. and M.A.A.A.-q.; formal analysis, M.E.C. and H.B.; investigation, M.E.C. and H.B.; resources, M.E.C., H.B.; data curation, M.E.C.; writing—original draft preparation, M.E.C. and H.B.; writing—review and editing, M.A.A.A.-q. and A.D.; visualization, M.E.C. and H.B.;

supervision, A.D.; project administration, A.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by National Natural Science Foundation of China (Grant No. 62150410434), and by LIESMARS Special Research Funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data used in this study are public datasets as mentioned in the main text.

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

References

1. Fan, H.; Du, W.; Dahou, A.; Ewees, A.A.; Yousri, D.; Elaziz, M.A.; Elsheikh, A.H.; Abualigah, L.; Al-qaness, M.A.A. Social Media Toxicity Classification Using Deep Learning: Real-World Application UK Brexit. *Electronics* **2021**, *10*, 1332. https://doi.org/10.3390/electronics10111332.
2. 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, 31 May 2019; pp. 187–196.
3. Aldjanabi, W.; Dahou, A.; Al-qaness, M.A.; Elaziz, M.A.; Helmi, A.M.; Damaševičius, R. Arabic Offensive and Hate Speech Detection Using a Cross-Corpora Multi-Task Learning Model. *Informatics* **2021**, *8*, 69.
4. Abdelgwad, M.M.; Soliman, T.H.A.; Taloba, A.I.; Farghaly, M.F. Arabic aspect based sentiment analysis using bidirectional GRU based models. *J. King Saud-Univ.-Comput. Inf. Sci.* 2021, *in press*.
5. Al-Smadi, M.; Qawasmeh, O.; Talafha, B.; Quwaider, M. Human annotated arabic dataset of book reviews for aspect based sentiment analysis. In Proceedings of the 2015 3rd International Conference on Future Internet of Things and Cloud, Rome, Italy, 24–26 August 2015; pp. 726–730.
6. Obaidat, I.; Mohawesh, R.; Al-Ayyoub, M.; AL-Smadi, M.; Jararweh, Y. Enhancing the determination of aspect categories and their polarities in Arabic reviews using lexicon-based approaches. In Proceedings of the 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Amman, Jordan, 3–5 November 2015; pp. 1–6. https://doi.org/10.1109/AEECT.2015.7360595.
7. Adel, H.; Dahou, A.; Mabrouk, A.; Abd Elaziz, M.; Kayed, M.; El-Henawy, I.M.; Alshathri, S.; Amin Ali, A. Improving Crisis Events Detection Using DistilBERT with Hunger Games Search Algorithm. *Mathematics* **2022**, *10*, 447. https://doi.org/10.3390/math10030447.
8. Obied, Z.; Solyman, A.; Ullah, A.; Fat'hAlalim, A.; Alsayed, A. BERT Multilingual and Capsule Network for Arabic Sentiment Analysis. In Proceedings of the 2020 International Conference On Computer, Control, Electrical, And Electronics Engineering (ICCCEEE), Khartoum, Sudan, 26 February–1 March 2021; pp. 1–6.
9. Oueslati, O.; Cambria, E.; HajHmida, M.B.; Ounelli, H. A review of sentiment analysis research in Arabic language. *Future Gener. Comput. Syst.* **2020**, *112*, 408–430.
10. Hamada, S.; Marzouk, R.M. Developing a transfer-based system for Arabic Dialects translation. In *Intelligent Natural Language Processing: Trends and Applications*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 121–138.
11. Al-Ibrahim, R.; Duwairi, R.M. Neural machine translation from Jordanian Dialect to modern standard Arabic. In Proceedings of the 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 7–9 April 2020; pp. 173–178.
12. Torjmen, R.; Haddar, K. Translation system from Tunisian Dialect to Modern Standard Arabic. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e6788. https://doi.org/10.1002/cpe.6788.
13. Hegazi, M.O.; Al-Dossari, Y.; Al-Yahy, A.; Al-Sumari, A.; Hilal, A. Preprocessing Arabic text on social media. *Heliyon* **2021**, *7*, e06191.
14. Tachicart, R.; Bouzoubaa, K. Moroccan data-driven spelling normalization using character neural embedding. *Vietnam. J. Comput. Sci.* **2021**, *8*, 113–131.
15. Husain, F.; Uzuner, O. Investigating the Effect of Preprocessing Arabic Text on Offensive Language and Hate Speech Detection. *Trans. Asian-Low-Resour. Lang. Inf. Process.* **2022**, *21*, 1–20.
16. Elnagar, A.; Yagi, S.; Nassif, A.B.; Shahin, I.; Salloum, S.A. Sentiment analysis in dialectal Arabic: a systematic review. In *International Conference on Advanced Machine Learning Technologies and Applications*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 407–417.
17. Xue, W.; Li, T. Aspect based sentiment analysis with gated convolutional networks. *arXiv* **2018**, arXiv:1805.07043.
18. Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S.; Al-Smadi, M.; Al-Ayyoub, M.; Zhao, Y.; Qin, B.; De Clercq, O.; et al. Semeval-2016 Task 5: Aspect basEd Sentiment Analysis. In *International Workshop on Semantic Evaluation*; Association for Computational Linguistics, San Diego, CA, USA, 2016; pp. 19–30.
19. Liu, N.; Shen, B. Aspect-based sentiment analysis with gated alternate neural network. *Knowl.-Based Syst.* **2020**, *188*, 105010.

20. Li, X.; Bing, L.; Zhang, W.; Lam, W. Exploiting BERT for End-to-End Aspect-based Sentiment Analysis. *arXiv* **2019**, arXiv:1910.00883. https://doi.org/10.48550/ARXIV.1910.00883.
21. Li, X.; Bing, L.; Li, P.; Lam, W. A unified model for opinion target extraction and target sentiment prediction. *AAAI Conf. Artif. Intell.* **2019**, *33*, 6714–6721.
22. Xu, B.; Wang, X.; Yang, B.; Kang, Z. Target embedding and position attention with lstm for aspect based sentiment analysis. In Proceedings of the 2020 5th International Conference on Mathematics and Artificial Intelligence, Chengdu, China, 10–13 April 2020; pp. 93–97.
23. 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.
24. Abdelgwad, M.M. Arabic aspect based sentiment analysis using BERT. *arXiv* **2021**, arXiv:2107.13290.
25. Al-Sarhan, H.; Al-So'ud, M.; Al-Smadi, M.; Al-Ayyoub, M.; Jararweh, Y. Framework for affective news analysis of arabic news: 2014 gaza attacks case study. In Proceedings of the 2016 7th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 5–7 April 2016; pp. 327–332.
26. Ashi, M.M.; Siddiqui, M.A.; Nadeem, F. Pre-trained word embeddings for Arabic aspect-based sentiment analysis of airline tweets. In *International Conference on Advanced Intelligent Systems and Informatics*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 241–251.
27. Al-Dabet, S.; Tedmori, S.; Mohammad, A.S. Enhancing Arabic aspect-based sentiment analysis using deep learning models. *Comput. Speech Lang.* **2021**, *69*, 101224.
28. Mohammad, A.S.; Hammad, M.M.; Sa'ad, A.; Saja, A.T.; Cambria, E. Gated Recurrent Unit with Multilingual Universal Sentence Encoder for Arabic Aspect-Based Sentiment Analysis. *Knowl.-Based Syst.* **2021**, 107540. https://doi.org/10.1016/j.knosys.2021.107540.
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.
30. 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.
31. Javaloy, A.; García-Mateos, G. Text normalization using encoder–decoder networks based on the causal feature extractor. *Appl. Sci.* **2020**, *10*, 4551.
32. Lourentzou, I.; Manghnani, K.; Zhai, C. Adapting sequence to sequence models for text normalization in social media. *Int. AAAI Conf. Web Soc. Media* **2019**, *13*, 335–345.
33. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
34. Sun, C.; Huang, L.; Qiu, X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. *arXiv* **2019**, arXiv:1903.09588.
35. Meftouh, K.; Harrat, S.; Jamoussi, S.; Abbas, M.; Smaili, K. Machine translation experiments on PADIC: A parallel Arabic dialect corpus. In Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, Shanghai, China, 30 October–1 November 2015; pp. 26–34.
36. Bouamor, H.; Habash, N.; Salameh, M.; Zaghouani, W.; Rambow, O.; Abdulrahim, D.; Obeid, O.; Khalifa, S.; Eryani, F.; Erdmann, A.; et al. The MADAR Arabic Dialect Corpus and Lexicon. In Proceedings of the LREC, Miyazaki, Japan, 7–12 May 2018.
37. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
38. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv* **2019**, arXiv:1910.03771.
39. 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, Spain, 12–13 December 2020; pp. 2054–2059.
40. Antoun, W.; Baly, F.; Hajj, H. Arabert: Transformer-based model for arabic language understanding. *arXiv* **2020**, arXiv:2003.00104.
41. Obeid, O.; Zalmout, N.; Khalifa, S.; Taji, D.; Oudah, M.; Alhafni, B.; Inoue, G.; Eryani, F.; Erdmann, A.; Habash, N. CAMeL tools: An open source python toolkit for Arabic natural language processing. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 7022–7032.
42. Ruder, S.; Ghaffari, P.; Breslin, J.G. INSIGHT-1 at SemEval-2016 Task 5: Deep Learning for Multilingual Aspect-based Sentiment Analysis. *arXiv* **2016**, arXiv:1609.02748. https://doi.org/10.48550/ARXIV.1609.02748.