applied sciences

MDPI

Article

Affect Analysis in Arabic Text: Further Pre-Training Language Models for Sentiment and Emotion

Wafa Alshehri [1,2,3,*], Nora Al-Twairesh [1,4] and Abdulrahman Alothaim [1,2]

[1] STC's Artificial Intelligence Chair, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia; twairesh@ksu.edu.sa (N.A.-T.); othaim@ksu.edu.sa (A.A.)
[2] Department of Information Systems, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia
[3] Department of Computer Sciences, College of Science and Arts, King Khalid University, Almajarda 63931, Saudi Arabia
[4] Department of Information Technology, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia
* Correspondence: waalshehri@kku.edu.sa

Abstract: One of the main tasks in the field of natural language processing (NLP) is the analysis of affective states (sentiment and emotional) based on written text, and attempts have improved dramatically in recent years. However, in studies on the Arabic language, machine learning or deep learning algorithms were utilised to analyse sentiment and emotion more often than current pre-trained language models. Additionally, further pre-training the language model on specific tasks (i.e., within-task and cross-task adaptation) has not yet been investigated for Arabic in general, and for the sentiment and emotion task in particular. In this paper, we adapt a BERT-based Arabic pretrained language model for the sentiment and emotion tasks by further pre-training it on a sentiment and emotion corpus. Hence, we developed five new Arabic models: QST, QSR, QSRT, QE3, and QE6. Five sentiment and two emotion datasets spanning both small- and large-resource settings were used to evaluate the developed models. The adaptation approaches significantly enhanced the performance of seven Arabic sentiment and emotion datasets. The developed models showed excellent improvements over the sentiment and emotion datasets, which ranged from 0.15–4.71%.

Keywords: sentiment analysis; emotion detection; pretrained language models; model adaptation; task-adaptation approach

Citation: Alshehri, W.; Al-Twairesh, N.; Alothaim, A. Affect Analysis in Arabic Text: Further Pre-Training Language Models for Sentiment and Emotion. *Appl. Sci.* **2023**, *13*, 5609. https://doi.org/10.3390/app13095609

Academic Editors: Pawel Dybala, Rafal Rzepka and Michal Ptaszynski

Received: 23 March 2023
Revised: 19 April 2023
Accepted: 28 April 2023
Published: 1 May 2023

1. Introduction

Natural language processing (NLP) is a field that is concerned with understanding, processing, and analysing natural languages (i.e., human languages). The evolution of the approaches used in NLP tasks is worth noting. Initially, the rule-based approach was dominant in the NLP field, which neglects to consider the contextual meaning of words, and which finds it difficult to cover all the morphologies of the language. With the growth in the availability and accessibility of data, the so-called machine-learning approach emerged. This method has a benefit in terms of accuracy when compared to the rule-based approach. It uses machine learning algorithms, but one of its drawbacks is that it requires complex manual feature engineering. With the emergence of neural networks and, more recently, deep learning, feature engineering has become automatic through the use of word embedding techniques, including Word2Vec [1], Glove [2], FastText [3], and others. Word vectors have been used in the NLP field, and they have achieved state-of-the-art results. Word embeddings are used to map all words into vectors of numbers in the vector space. Language models use pre-trained word embedding as an additional feature to initiliase the first layer of the basic model. The limitations of the word embeddings models are that they cannot handle out-of-vocabulary (OOV) words and meaning or context-dependent representations are lacking.

However, for model training, machine learning and deep learning approaches demand extensive amounts of labelled data, which is time-consuming to annotate and prepare. At present, significant evolution is noticeable in the NLP field, particularly with the emergence of transfer learning. This has reduced the need for massive amounts of training examples. In many NLP applications, most of the recent research that has applied transfer learning techniques has been associated with state-of-the-art results. Transfer learning, according to [4], "The improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned". Transfer learning with pre-trained language models has attracted the interest of the research community in recent years, thus relying on the so-called semi-supervised approach. The language model trains in an unsupervised manner with a significant amount of unlabeled data (corpora), followed by a supervised process of fine-tuning the language model with a labelled dataset that is small and task-specific.

With the advancement of technology and the widespread nature of social networking sites, people have become more expressive of their sentiment and emotion, and others' opinions may also influence them. Many entities have started to consider customer opinions relating to their products or services. The Arabic language is one of the most popular languages in the world and was ranked fourth among the languages used on the internet [5], and it is also the primary language for 22 countries [6]. Therefore, there is a great need for tools and models to analyse Arabic sentiment and emotions on specific topics, phenomena, or trends, which can be benefited from several fields. Affect is the superordinate group; emotions and sentiments are statuses within this group. In other words, affect is a general term which includes both emotions and sentiment states. Sentiment analysis and emotion detection are both NLP tasks that have emerged as hot topics of interest in the NLP research community. According to the Oxford Dictionary [7], a sentiment refers to "a feeling or an opinion, especially one based on emotions", whereas an emotion is "a strong feeling such as love, fear, or anger; the part of a person's character that consists of feelings.", where we can infer that sentiment is a general interpretation of emotion. Sentiment is classified as positive, negative, or neutral, or—using an expanded scale—as very positive, positive, neutral, negative, or very negative. Emotions are often classified according to well-known models, including the Ekman model [8], into happiness, sadness, fear, anger, surprise, and disgust, or using the Plutchik wheel of emotion [9].

Emotion and sentiment analysis in text depends essentially on the language used. This study aims to analyse sentiment and emotion in Arabic, which is one of the most challenging languages. The Arabic language has several varieties, including classical Arabic, modern standard Arabic (MSA), and other dialects. Many challenges are facing the field of text emotion and sentiment analysis in the Arabic language in particular, including the lack of resources, the diversity of dialects that have no standard rules, and the detection of the implicit expression of sentiment or emotion. Furthermore, one root word may be written in more than one form, or one word may have more than one meaning. Additionally, the diacritics change the meaning of the words [5]. The Arabic language differs from other languages due to its morphological richness and complex syntactic synthesis. Therefore, NLP tasks for text emotion and sentiment analysis become more complex in the Arabic language. These challenges, along with others, have delayed progress in this research area, meaning that these tasks have not been adequately investigated and explored in Arabic compared to English. In Arabic, a degree of progress has been recorded in the field of sentiment analysis, where sentiments are typically classified as positive, negative, or neutral. However, progress in emotion detection task is ongoing, and few studies have been conducted that classify emotions deeply (e.g., emotions classification according to Ekman [8], or Plutchik [9]). The evolution that occurred in this area, especially the exploitation of transfer learning and advanced pre-trained language models, led to overcoming many of this field challenges, as well as substantial performance improvements. Arabic research papers predominantly employ machine learning or deep learning algorithms, as opposed to pre-trained language models.

Recent efforts in these fields have focused on adapting pre-trained language models to specific domains and tasks using domain-specific or task-specific unlabeled corpora, by the continuation of the pre-training of language models on this task or domain. Using either a domain-adaptation approach [10–12] or a task-adaptation approach [13,14], model adaptation has led to significant performance enhancements in the English language. As far as we know, model adaptation approaches, especially additional pre-training of the language model on a specific domain, have only been used in two Arabic language studies [15,16]. However, classifying sentiment and emotion was not the focus of these studies. Additionally, further pre-training the language model on a specific task (i.e., within-task and cross-task adaptation) has not been investigated for Arabic in general and for sentiment and emotion tasks in particular. This study aims to tackle these problems and fill these gaps by developing models with the overall aim of advancing the current state of sentiment and emotion classification tasks for Arabic. The pre-trained language model QARiB [17], which has achieved state-of-the-art results in several NLP tasks, including sentiment analysis and emotion detection, was used in this study. QARiB is further pre-trained using sentiment and emotion-specific datasets, assuming that the small task-specific datasets given during the pre-training process are sufficient to improve model performance in that task. The developed model was then evaluated by fine-tuning it on seven sentiment and emotion datasets. In particular, the contributions of this study are:

- Develop five new Arabic language models: QST, QSR, QSRT, QE3, and QE6, which are the first task-specific adapted language models based on QARiB, for Arabic sentiment analysis and emotion detection tasks. The developed models significantly enhanced the performance of seven Arabic sentiment and emotion datasets, and the research community can use these models for sentiment and emotion tasks;
- Conduct comprehensive experiments to investigate the impact of the within-task and cross-task adaptation approaches on the performance of sentiment and emotion classification;
- Analyse the influence of the genre of training datasets (i.e., tweets and reviews) utilised for model adaption on the performance of sentiment classification;
- Make the newly adapted models available to the public (https://huggingface.co/NLP-EXP, accessed on 1 March 2023).

The remainder of the paper is organised as follows: Section 2 offers a concise literature review on the classification of sentiment and emotion in Arabic text. In Section 3, the approach proposed for developing the models, pre-training datasets, and all necessary pre-processing steps and tokenisation is described. In Section 4, the experimental setup is described, including the evaluation datasets, the baseline model, the fine-tuning architecture, and the hyperparameter choices for fine-tuning our models. The results of the experiment are presented and discussed in Section 5. Section 6 concludes the paper besides outlining a few other future directions.

2. Literature Review

This section reviews the literature on the classification of sentiment and emotion in Arabic text using various approaches, including lexicon-based, machine learning, deep learning, and fine-tuning Transformer-based language models.

2.1. Sentiment Analysis in Arabic

Recently, the task of sentiment analysis has attracted attention in the Arabic NLP community. Before the emergence of transfer learning, a single trained model was used for a single task. Building a model from scratch is costly in terms of training time, memory, materials, and the data required to train the model. In particular, deep learning models require large amounts of labelled data to train models and generate satisfactory results. In fact, generating usable datasets requires time and effort due to the limited availability of data not only in general but also in the Arabic language. Therefore, the advent of transfer learning enabled many of these problems to be overcome. It became possible to use a single model for multiple tasks instead of building a specific model for each task. Many models

using transfer learning have achieved state-of-the-art results across different NLP tasks. In several languages, pre-trained language models have yielded great results in sentiment analysis [18–20]. To achieve similarly high outcomes in Arabic, several works employed different Arabic Transformer-based pre-trained language models for sentiment analysis.

The Transformer architecture, which was introduced in [21], is entirely dependent on the attention mechanism, which includes the encoder and decoder parts. The BERT (bi-directional encoder representations from Transformers) language model [18] is based on the Transformer architecture and is composed of a set of Transformer encoders layered on top of each other. Two objective functions were used to train BERT. The first being the masked language modelling (MLM) objectives, which use the special <MASK> token to randomly mask samples of input tokens in order to predict the word given their context. BERT was also trained on the next sentence prediction (NSP) objective. Given two sentences, the model predicts whether or not they follow each other.

AraBERT, as proposed by Antoun et al. [22], was the first Transformer-based pre-trained language model in Arabic based on the BERT model. AraBERT was pre-trained on a ~24 GB Arabic corpora and fine-tuned for three NLP tasks, one of which was sentiment analysis. The authors fine-tuned and evaluated AraBERT for sentiment classification on five sentiment datasets, and the model achieved state-of-the-art results compared to mBERT [18] and hULMonA [23]. Several efforts [24,25] have employed the AraBERT model for sentiment classification.

Further, numerous approaches have achieved high performance in this field by fine-tuning the Arabic Transformer-based pre-trained models for sentiment analysis, such as mBERT [18], AraELECTRA [26], ARBERT and MARBERT [27], XLM-R [28], QARiB [17], CAMeLBERT [29], GigaBERT [16], DziriBERT [30], AraXLNet [31], Arabic-ALBERT (https://ai.ku.edu.tr/arabic-albert/, accessed on 1 February 2022), and ArabicBERT [32]. In [33], Elmadany et al. introduced ORCA, a publicly accessible benchmark for evaluating Arabic language understanding tasks. ORCA covers a variety of Arabic varieties and a range of challenging tasks using 60 datasets across seven Natural Language Understanding (NLU) task clusters. The task clusters include (1) sentence classification, (2) topic classification, (3) structured prediction, (4) semantic similarity, (5) natural language inference, (6) question-answering, and (7) word sense disambiguation. The authors used ORCA to compare 18 multilingual and Arabic pre-trained language models and created a public leaderboard with a unified evaluation metric (ORCA score) to support future research. ORCA includes sentiment analysis as one of its tasks; 19 available datasets were used to construct this task. The best performance on this task was attained by the AraElectra language model, which achieved 80.86%. As for the dialect tasks in ORCA, MARBERTv2 [27] and QARiB [17] get the highest ORCA scores, respectively.

At present, research in these fields has been directed towards adapting pre-trained language models to specific domains and tasks using domain-specific or task-specific unlabelled corpora. Model adaptation techniques have given rise to large performance gains in the English language. For example, using a domain-adaptation approach in [10–12], using the task-adaptation approach in [13,14]. In the biomedical domain, for example, [10] developed BioBERT, a BERT model that was further pre-trained on biomedical corpora, and evaluated on various biomedical text mining tasks, such as question answering (QA), named entity recognition (NER), and relation extraction (RE). Further, clinical BERT was introduced in [11] by continuing the pre-training of BERT and BioBERT models using clinical notes. Models were trained for 150k steps and fine-tuned on five NER and natural language inference (NLI) datasets. The results indicate that pre-training language models on biomedical and clinical corpora facilitate the comprehension of complex medical texts. In order to undertake a variety of NLP tasks in the field of finance, Araci et al. [12] developed FinBERT, a BERT-based language model. The model is additionally pre-trained using the financial corpus TRC2-financial, which contains approximately 400k sentences. The model was evaluated using two sentiment datasets: Financial PhraseBank and FiQA

Sentiment. Experiments indicate that FinBERT obtained state-of-the-art results on both datasets, improving accuracy by 15%.

In addition, [13] attempted to adapt the RoBERTa language model [34] using in-domain, within-task, and cross-task model adaptation approaches. For domain-adaptation, the authors additionally pre-trained RoBERTa for 12.5k steps in the biomedical (BioMed), computer science (CS), news, and reviews domains. For all domains, they evaluate each language model using two text classification datasets: CHEMPROT and RCT for BioMed, ACL-ARC and SCIERC for CS, HYPERPARTISAN and AGNEWS for News, and HELPFULNESS and IMDB for Reviews. The domain adaptation approach exhibits performance enhancements over the RoBERTa model on all datasets, with the exception of the AGNEWS dataset. For task adaptation, the RoBERTa was additionally pre-trained on each of the aforementioned datasets for 100 epochs, followed by an evaluation on the same dataset to determine the efficacy of the task-adaptation approach. Compared to the domain-adaptation approach, the task-adaptation approach utilises fewer, but more task-relevant data and is cheaper to implement. The outcomes of this approach were comparable to those of domain adaptation, and improvements over RoBERTa were demonstrated for all datasets. Lastly, the authors conducted an experiment utilising an approach for cross-task transfer. For instance, the RoBERTa is further pre-trained using the HYPERPARTISAN dataset before being evaluated and fine-tuned using the AGNEWS dataset. While the task-adaptation approach has been shown to be effective, the cross-task approach has been shown to have negative effects. A study conducted by [14] also demonstrated the effectiveness of domain and task adaptation approaches. Where BERT was additionally pre-trained using seven text classification datasets, including IMDB, Yelp P., Yelp F., TREC, Yahoo! Answers, AG's News, and DBPedia, which cover three domains: sentiment, topic, and question. The adapted models were subsequently evaluated using the aforementioned datasets, as additional pre-training is contributing to improving the performance of BERT for a particular task.

The domain-adaptation approach was used in Arabic NLP research by [15], they developed a language model in the COVID-19 domain, by further pre-training AraBERT and mBERT using around 1 million tweets. The models evaluated on the ARACOVID19-MFH dataset cover different NLP tasks, such as fake news detection, opinion mining, hate speech, etc. Another work by [16], was introducing a domain-specific language model pre-trained on large-scale news corpora. They utilised roughly 13 million news articles from the Gigaword corpus to further pre-train the XML-RoBERTa model. The model evaluated four NLP tasks: named entity recognition (NER), part-of-speech (POS), relation extraction (RE), and argument role labelling (ARL). Other works that utilised domain adaptive approaches for Arabic sentiment analysis are found in [35–37]. The results of these works show there is a significant performance improvement that could result from domain-specific adaptation.

2.2. Emotion Detection in Arabic

Notably, limited studies have involved the detection of emotion in Arabic text. Most prior studies have used traditional methods, such as the lexicon-based approach [38]. In addition, machine learning or deep learning algorithms and the AIT dataset (affect in tweets for SemEval-2018 competition Task 1) have been utilised in the majority of Arabic emotion detection studies [39–42], due to the scarcity of Arabic resources for this task.

Limited studies have fine-tined Arabic Transformer-based models for emotion detection. One of the studies on emotion and sentiment showcased the AraNet toolkit by Abdul-Mageed et al. [43], where mBERT was fine-tuned on many of the available datasets for different tasks, including sentiment and emotion analysis. For the sentiment task, mBERT was fine-tuned and tested using 15 sentiment datasets, collectively containing approximately 126,766 examples. For the emotion task, mBERT was fine-tuned using two datasets, LAMA-DINA and LAMA-DIST, collectively containing approximately 189,903 tweets. To the best of our knowledge, this was the first research to use an Arabic Transformer-based

model (i.e., mBERT) for emotion detection. Notably, AraNet achieved state-of-the-art performance on these tasks.

Using the AraNet-Emo dataset [43], the developer of the ARBERT and MARBERT models [27] fine-tuned these models for emotion classification. In comparison to AraNet [43], XLM-R [28], and AraBERT [22], MARBERT obtained state-of-the-art results with an F1-score of 75.83%. Additionally, the QARiB language model [17] was fine-tuned for emotion detection using the AIT emotion classification (E-c) task dataset. The model attained state-of-the-art performance and outperformed the AraBERT [22], mBERT [18] and ArabicBERT [32] with a macro-F1 score of 46.8%. Elfaik et al. in [44] used AraBERT for extracting features and an attentional LSTM-BiLSTM model for emotion classification. Utilising the AIT dataset, exhaustive experiments were conducted, in which the proposed approach performs better than many versions of the mBERT and AraBERT models. The authors of [45] suggested an ensemble deep-learning method for detecting emotion in Arabic Tweets. The AIT dataset [39] was evaluated using three deep learning models individually (Bi-LSTM, Bi-Directional gated recurrent unit Bi-GRU and MARBERT), and compared to the developed ensemble model. The developed ensemble model significantly outperforms the individual models, as shown by the increase in macro F1 score varying from 5.3% to 23.3%.

Using the AIT dataset [39], Al-Twairesh [46] conducted an experimental study on the development of language models. Extensive experiments were carried out to examine the effectiveness of several language models (including the traditional TF-IDF, different versions of AraVec [47], AraBERT [22], and ArabicBERT [32] models, and multi-DialectBert [48]) on the emotion detection task. The results demonstrate that the ArabicBERT-Large model showed the best performance. One of the most recent works was proposed by Mahmoud et al. [49]. Researchers released the "Arabic Egyptian COVID-19 Twitter Dataset (ArECTD)", one of the largest Arabic emotion detection datasets. The dataset included roughly 78k tweets that were classified into ten emotion classes: "sadness", "fear", "sarcasm", "sympathy", "anger", "surprise", "love", "joy", "hope", and "none." The dataset was evaluated using two Arabic language models, AraBERT and MARBERT, achieving accuracy values of 70.01% and 72.5%, respectively.

To summarise, new pre-trained language models, have shown significant developments in sentiment and emotion classification. However, gaps exist due to insufficient research activity in this area for the Arabic language. In Arabic, most research papers in NLP have focused on using machine learning or deep learning algorithms to address sentiment and emotion classification problems. The evolution that has occurred in the NLP area, especially in the exploitation of transfer learning and advanced pre-trained language models (Transformer-based models), has not been significantly investigated in the Arabic language. However, compared to English, in Arabic studies, machine learning or deep learning algorithms were utilised to analyse sentiment and emotion more often than current pre-trained language models. In particular, the Arabic emotion detection task has limited studies compared to the Arabic sentiment analysis task. In addition, most of the studies have used a limited number of datasets (i.e., experiments using one or two datasets), or small datasets; for example, in emotion detection studies, most have used the AIT dataset. The reason could be the limited number of resources available for Arabic sentiment and emotion classification tasks. Furthermore, model adaptation approaches have given rise to large performance gains in the English language, whether using domain-adaptation approaches [10–12] or task-adaptation approaches [13,14]. Adapting the pre-trained language model to a specific domain or task means continuing the pre-training of language models on this task or domain using an unlabelled domain-specific or task-specific dataset. To the best of our knowledge, model adaptation approaches, in particular, further pre-training the language model on a specific domain have only been undertaken in two studies for the Arabic language [15,16], whereas classifying sentiment and emotion were not the focus of these studies. Additionally, further pre-training the language model on a specific task (i.e., within-task and cross-task adaptation) has not yet been investigated for Arabic in

general, and for the sentiment and emotion task in particular. Nevertheless, amid the emerging advancements in Arabic sentiment and emotion classification, further study and experimentation are still needed to address the existing gaps and enhance performance in this field.

3. Methodology

The current study aims to develop pre-trained language models to boost the current state of Arabic sentiment analysis and emotion detection tasks. This section provides more details about the adaptation approach utilised in this work and the developed pre-trained language models. Furthermore, the pre-taring data collection, as well as all required pre-processing steps and tokenisation, are described. Finally, the section concludes with a presentation of the model's evaluation metrics.

3.1. Task-Adaptation Approach

Recent research has demonstrated that further unsupervised pre-training of the language model on the task-specific dataset, followed by fine-tuning on the supervised target task dataset, can yield substantially better performance than directly supervised target task fine-tuning [14]. Further pre-training enables continued training of the pre-trained language model on domain-specific or task-specific corpora instead of building or pre-training the model from scratch, which is more time-consuming and has a high computational cost. For these reasons and to achieve these research goals, we use the task-adaptation approach to enhance and obtain a better result for Arabic sentiment and emotion classification tasks. Due to the specialised language used in the emotion and sentiment context, general-purpose models are inadequate. The focus of this work is on NLP transfer learning or model adaptation methodologies that appear to be a promising solution to this challenge. Moreover, due to the scarcity of Arabic resources for sentiment and emotion tasks, we believe that pre-trained language models can help to solve these problems since they require fewer annotated samples and can be further pre-trained using task-specific (sentiment and emotion) corpora. Since the distribution of the text for the target task differed from general corpora, we used QARiB pre-trained language model [17], to develop models that could learn the semantic relations in the target task's text.

The QARiB model (QCRI Arabic and Dialectal BERT) [17] uses a BERT-based architecture [18], both of which are Transformers-based architectures. QARiB [17] is trained using only the MLM objective with a masking probability of 15%, and the NSP objective was excluded. QARiB has five versions that differ in the size of training datasets, the mixture of formal (MSA) and informal (dialect) Arabic text and using a Farasa [50] tokeniser or not. Arabic Gigaword Fourth Edition [51], Abulkhair Arabic Corpus [52], Open Subtitles [53], and a collection of Arabic tweets constitute the training dataset. A Byte-Pair-Encoding (BPE) tokeniser [54] was employed for dataset tokenisation with a vocabulary size of 64k. The architecture of all of these models is the same as the BERT-base model, with 12 encoder layers, a hidden size of 768, and 12 multi-head attention heads. A BERT-base-QaRiB model version was utilised in our study, and we further pre-trained the model using task-specific datasets. QARiB has achieved excellent results on a variety of Arabic NLP tasks, including sentiment analysis, named entity recognition, dialect identification, emotion classification, and offensive detection, despite employing almost the same structure across tasks. Comparing the model against various Arabic pre-trained language models, including AraBERT [22], mBERT [18], and ArabicBERT [32], showed state-of-the-art results. Adapting QARiB for sentiment and emotion tasks could potentially aid or benefit numerous NLP studies on sentiment and emotion and boost task outcomes. This can be accomplished by investigating whether the newly adapted language models can provide better outcomes than the QARiB language model on these tasks. To determine if this adaptation would be beneficial for sentiment and emotion classification tasks, two model adaptation approaches were implemented as follows: first, within-task adaptation; second, cross-task adaptation. More details about this process are given in the following subsections.

3.1.1. Pre-Training Datasets Collection

The initial phase of this method, as shown in Figure 1, is to gather pre-training task-specific datasets. The further pre-training of the language model on particular tasks needs task-specific unlabelled data. Building or generating a new dataset requires much time and effort. For this reason, we gathered further pre-training datasets from existing and available Arabic sentiment and emotion datasets. We built three sentiment datasets and one emotion dataset by augmenting different available datasets. The datasets were constructed as follows:

1. Sentiment Tweets Dataset (STD): a collection of fourteen sentiment datasets, sourced from Twitter. The statistics for these datasets are reported in Table 1. It contains the following datasets: Arabic Jordanian General Tweets (AJGT) [55], Arabic Speech Act and Sentiment (ArSAS) [56], Arabic Sentiment Twitter Dataset for the Levantine Dialect (ArSenTD-LEV) [57], Arabic-Dialect [58], BBN-Dataset [59], Syrian-Dataset [59], The Arabic Tweets Sentiment Analysis Dataset (ATSAD) [60], The Arabic Sentiment Tweets Dataset (ASTD) [61], SemEval-2017 [62], ArSarcasm [63], The Tweets and Emojis Arabic Dataset for Sentiment Analysis (TEAD) [64], Affect in Tweets (AIT) [39], Multi-Domain Arabic Resources for Sentiment Analysis (MARSA) [65], and AraSenti-Tweet [66]; Table 1 illustrates a summary of the STD dataset statistics which contains over 682,000 tweets, approximately 445,395 positive tweets and approximately 238,355 negative tweets. We only kept the positive and negative classes in our experiment. Other classes such as neutral, mixed, or objective were eliminated. To obtain more specific or relevant sentiment data, we utilised the manually annotated sentiment datasets. In total, 11 out of the 14 datasets were manually annotated. The last column in Table 1 shows the average sequence length for each dataset, which varies from 10 to 23. It can be noticed that the STD has a sequence length of 13 words on average.
2. Sentiment Reviews Dataset (SRD): we augmented four reviews sentiment datasets to build this dataset. The review datasets were utilised as follows: The Opinion Corpus for Lebanese Arabic Reviews (OCLAR) [67], Large-scale Arabic Book Review (LABR) [68], Hotel Arabic-Reviews Dataset (HARD) [69], and the Book Reviews in Arabic Dataset (BRAD 1.0) [70]; The rating in reviews datasets is regarded as a human annotation, in which people score the services using stars between 1 and 5. Scores of 4 and 5 are regarded as positive, scores of 1 and 2 are regarded as negative, and a score of 3 is regarded as neutral. Accordingly, we kept the rating of stars 5 and 4 as a positive class and the rating of stars 2 and 1 as a negative class. Additionally, we regarded 3-star ratings to be neutral and eliminated them from our experiment. Table 2 illustrates a summary of the SRD Dataset Statistics, which contains over 751,000 reviews. About 616,700 of the reviews are positive, while 134,773 are negative. The dataset covers several domains, including restaurants, hotels, and books. The last column in Table 2 displays the average sequence length for each dataset, which ranges from 13 to 80. It is notable that the SRD has an average sequence length of 54 words. The nature of tweets differs from that of reviews in terms of sentence length. Tables 1 and 2 illustrate the average sequence lengths of the STD and SRD datasets, respectively. Overall, the SRD has the longest sentence with an average of 54 words, while the STD's average sequence length is 13. We created datasets with two different data types (tweets and reviews) to investigate if using different data types (tweets and reviews) in the further pre-training phase, would affect the final results.
3. Sentiment Tweets–Reviews Dataset (STRD): This dataset combines the STD and SRD datasets and consists of 1,433,657 sentences. Table 3 shows the dataset's statistics. It is notable that the STRD has an average sequence length of 35 words;
4. Emotion Tweets Dataset (ETD): a collection of five emotion datasets which contains the following: Emotional-Tone [71], LAMA-DINA [72], Affect in Tweets (AIT) [39], AraEmoTw [73], and The SemEval-2018 Affect in Tweets Distant Supervision Corpus [39]. We primarily used Ekman's emotional classes in our experiment, which

include joy, sadness, anger, disgust, fear, and surprise. Therefore, tweets annotated with any other classes were excluded. Table 4 illustrates a summary of the ETD Dataset statistics, which comprises roughly 1.2 million tweets. There are around 227,518 tweets for the joy class, 346,696 for sadness, 344,899 for fear, 291,500 for anger, 27,166 for surprise, and 24,431 for disgust. The average sequence length for each dataset is shown in the final row of Table 4, and it ranges from 14 to 32. The ETD has an average sequence length of 19 words.

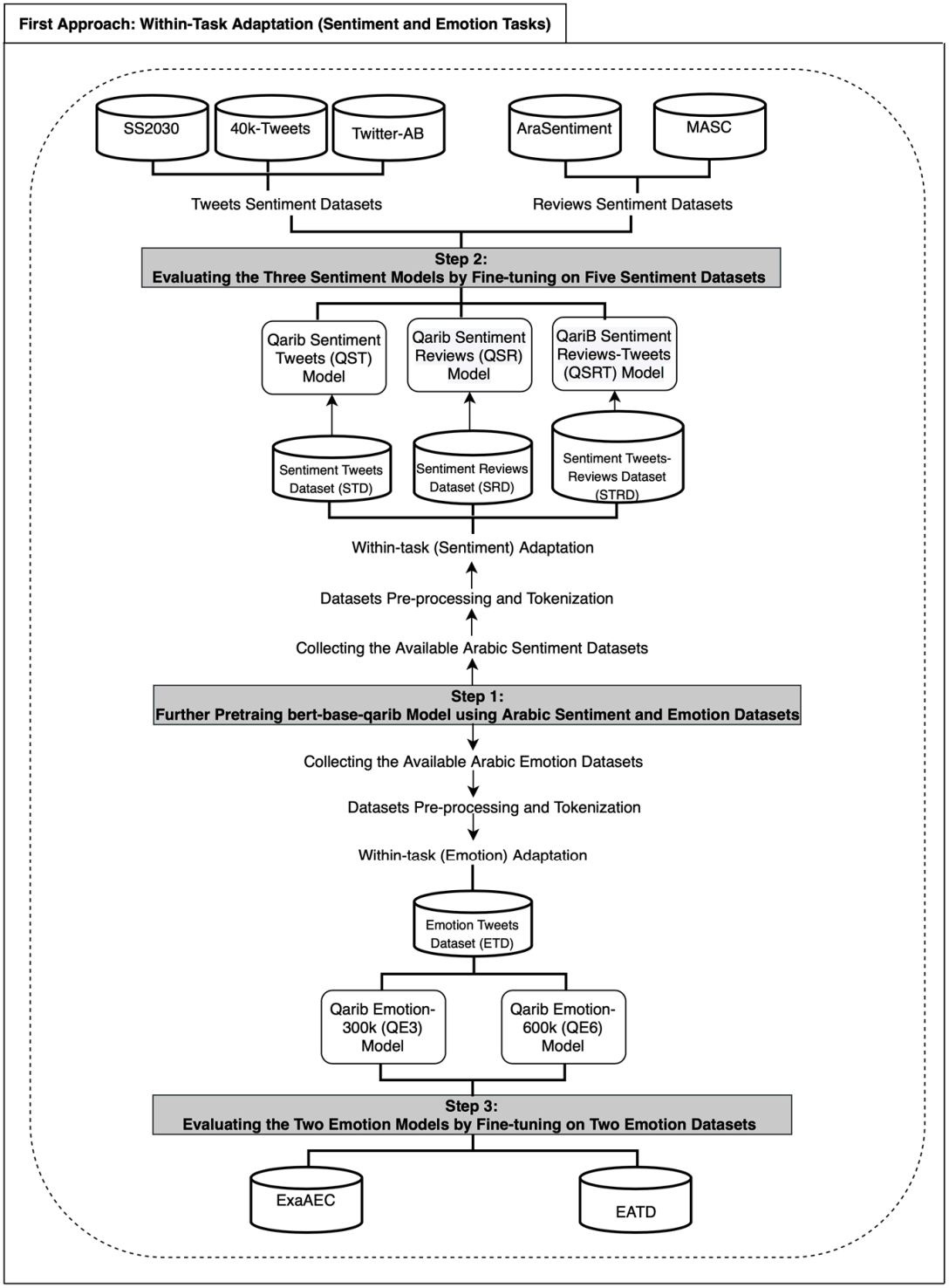

Figure 1. Language Models Development Process Using Within-Task Adaptation Approach.

Table 1. Sentiment Tweets Dataset (STD) Statistics.

Dataset-Name	Size	#Classes	#Pos	#Neg	#Other Classes	#Tweets ≥ 3	Dup	Size-after-Pre Processing	Avg SeqLen
AJGT [55]	1800	2	813	727	-	260	-	1540	10
ArSAS [56]	19,897	4	4323	7325	8113	79	57	11,648	23
ArSenTD-LEV [57]	4000	5	895	1292	885	0	928	2187	23
Arabic-Dialect [58]	52,210	3	5546	15,086	30,033	1118	427	20,632	21
BBN-Dataset [59]	1200	3	481	558	127	34	-	1039	10
Syrian-Dataset [59]	2000	3	447	1349	202	0	2	1796	19
ATSAD [60]	56,795	2	16,106	16,476	-	3980	20,233	32,582	13
ASTD [61]	9693	4	736	1592	7274	91	-	2328	17
ArSarcasm [63]	10,547	3	1634	3473	5340	71	29	5107	17
SemEval2017 [62]	3352	3	728	1108	1470	6	40	1836	17
TEAD [64]	555,923	2	391,530	162,763	-	673	957	554,293	13
Affect in Tweets [39]	1800	7	720	808	262	10	-	1528	17
MARSA [65]	56,782	3	16,647	19,858	18,726	384	1167	36,505	15
AraSenti-Tweet [66]	17,573	4	4789	5940	6461	367	13	10,729	16
Total (STD)	683,750	2	445,395	238,355	78,893	7073	25,406	682,197	13

Table 2. Sentiment Reviews Dataset (SRD) Statistics.

Dataset-Name	Size	#Positive	#Negative	#Neutral	#Tweets ≥ 3	Dup	Size-after-Preprocessing	Avg Len of Seq
OCLAR [67]	3916	1595	275	418	1587	41	1870	13
LABR [68]	63,257	39,069	7568	12,201	2328	2091	46,637	66
HARD [69]	409,562	263,453	51,864	80,326	8570	5349	315,317	22
BRAD [70]	510,599	312,583	75,066	106,785	14,429	1736	387,649	80
Total (SRD)	751,473	616,700	134,773	199,730	26,914	9217	751,473	54

Table 3. Sentiment Tweets-Reviews Dataset (STRD) Statistics.

Dataset-Name	Size	#Positive	#Negative	Duplicate	Avg Len of Seq	Size-after-Preprocessing
STD	682,197	445,395	238,355	-	13	-
SRD	751,473	616,700	134,773	-	54	-
Total (STRD)	1,433,670	1,062,095	373,128	13	35	1,433,657

Table 4. Emotion Tweets Dataset (ETD) Statistics.

Dataset-Name	Emotional-Tone [71]	LAMA-DINA [72]	Affect in Tweets [39]	AraEmoTw [73]	SemEval-2018 AIT DISC [39]	Total (ETD)
Size	10,065	8502	5600	226,774	1,019,435	1,262,210
#Joy	1169	1265	1389	38,591	185,104	227,518
#Sad	1222	964	792	23,871	319,847	346,696
#Fear	1140	1376	797	44,036	297,550	344,899
#Anger	1423	902	1390	70,851	216,934	291,500
#Surprise	992	1141	-	25,033	-	27,166
#Disgust	-	986	-	23,445	-	24,431
#Other-Classes	3832	1777	-	-	-	5609
#Tweets ≥ 3	279	88	37	421	-	825
#Duplicate	8	3	1195	526	-	1737
Size-after-preprocessing	5946	6634	4368	225,827	1,019,435	1,262,205
Avg Len of Seq	14	15	17	32	16	19

3.1.2. Pre-Training Datasets Pre-Processing

The datasets collected in the previous sections were already pre-processed by their authors. Therefore, we performed light pre-processing to prepare the datasets for the further pre-training task. Tables 1–4 show the statistics of the STD, SRD, STRD, and ETD datasets. The following pre-processing steps were performed:

- The URLs, user mentions, and hashtags present in any of the collected sentences were replaced with the tokens [رابط,مستخدم] and [هاشتاق];
- Sentences with three words or fewer were removed;
- Null and duplicated were eliminated.

3.1.3. Pre-Training Datasets Tokenisation

The original QARiB model [17] utilises byte-pair-encoding (BPE) [54] tokenisation with a set of vocabulary generated from several Arabic corpora. Following [17], we utilised a BPE tokeniser with a 64k tokens vocabulary to tokenise the dataset in our experiment. As stated in [10], employing a custom vocabulary (e.g., domain-specific vocabulary) prevents benefiting from the pre-training from the BERT checkpoint. Moreover, in [74], it was shown that pre-training with a custom vocabulary has outcomes that coincide with the outcomes that result from pre-training with a general vocabulary. Due to these reasons, our models have been further pre-trained with the QARiB model's BPE vocabulary.

3.1.4. Within-Task Adaptation Approach

The main phase of the approach presented in Figure 1 is to continue pre-training QARiB with sentiment and emotion datasets that we collected and preprocessed in the previous sections. The proposed approach (i.e., within-task adaptation) can be explained as: The QARiB language model is further pre-trained using training data for a target task. The model is trained on an unlabelled task-specific dataset (e.g., sentiment dataset) and evaluated by performing fine-tuning on the labelled dataset from the same task (e.g., sentiment dataset). This can be expressed as:

$$QARiB_T \rightarrow PretrainingDataset_T \rightarrow FinetuningDataset_T \quad (1)$$

"T" refers to the target task, which might be either sentiment or emotion. The overall process of the within-task adaptation approach of the model's development is illustrated in Figure 1.

The adaption of our models followed the settings for training the QARiB language model since this model was trained using only the masked language modelling objective. In this experiment, we further pre-trained five versions of the QARiB language model from the BERT-base-QaRiB checkpoint. Three of them were trained with sentiment-specific datasets including STD, SRD, and STRD. The objective was to investigate the impact of continuous pre-training on sentiment task performance. Furthermore, two models were trained using ETD, which is an emotion-specific dataset. The goal is to determine whether further pre-training may improve the results of emotion classification. The four datasets were divided into train and test sets at a ratio of 80% and 20%, respectively. We utilised the training script run_mlm.py provided by huggingface (https://github.com/huggingface/transformers.git, accessed on 1 February 2022), and we ran it on a single GPU provided by Google Colab Pro+. Every 5000 steps, a model checkpoint was saved. The models were trained using the QARiB model's default hyperparameters, as stated below:

- A batch size of 64;
- A maximum sequence length of 64 words;
- A learning rate of 8×10^{-5}.

The next subsections offer a detailed explanation of each model we developed for each task (sentiment and emotion).

Sentiment Models

For the sentiment task, we further pre-train three different sentiment models using the three sentiment datasets: STD, SRD, and STRD. The first model is the QARiB-Sentiment-Tweets (QST) Model, which is a QARiB language model that was further pre-trained for 300 k training steps and roughly 28 epochs using the STD (shown in Table 1). The second model is the QARiB-sentiment-reviews (QSR) Model, which is a QARiB language model that was further pre-trained for 300 k training steps and roughly 26 epochs using the SRD

(shown in Table 2). Despite the fact that the QST and QSR models were trained for the same number of training steps (300 k steps) and roughly the same data size (see Tables 1 and 2), Table 5 shows that the QSR has lower validation loss and better perplexity results than the QST. The reason could be that the review data has a longer text and is less noisy than the tweet dataset. As a result, we decided to continue training from the last checkpoint of the QSR model and trained the third model, which we call the QARiB-sentiment-reviews-tweets (QSRT) model, for an additional 300 k training steps (600 k in total with QSR training steps). The QSRT model was trained for roughly 52 epochs using STRD. Table 5 shows that the QSRT model outperformed the QST and QSR models in terms of training results. This was expected, given the increased amount of training data and the variety of data types (i.e., tweets and reviews). Table 5 provides additional details about the model's training results.

Table 5. The Adapted Model's Training Results.

Model	Training Steps	Epochs	Train_loss	Val_loss	Perplexity
QST Model	300 k	28.46	0.4839	2.704	14.9399
QSR Model	300 k	26.08	0.7821	2.6746	14.5071
QSRT Model	600 k	52.16	0.0209	2.5515	12.8269
QE3 Model	300 k	15.52	0.467	2.7762	16.0576
QE6 Model	600 k	31.04	0.435	2.7778	16.0838

Emotion Models

For the emotion task, we further pre-train two different emotion models using the emotion dataset ETD. The first model is the QARiB-emotion-300k (QE3) model, which is a QARiB language model that was further pre-trained for 300 k training steps and roughly 28 epochs using the ETD (shown in Table 4). The second emotion model, known as the QARiB-Emotion-600k (QE6) Model, was trained by continuing the training from the last checkpoint of the QE3 model using the same dataset ETD. The QE6 model trained for an additional 300 k steps (600 k in total including the QE3 training steps) and around 31 epochs. Table 5 shows that there was no improvement in the training results of the QE6 models when compared to the QE3 models. In fact, there was an increase in QE6 validation loss and perplexity. According to Table 5, the QSRT model has the lowest validation loss and perplexity score. Furthermore, we can see that the sentiment models outperform the emotion models in terms of validation loss and perplexity score. Afterwards, each of the three sentiment models was evaluated by fine-tuning using various sentiment datasets. Additionally, the emotion models were tested by fine-tuning them using various emotion datasets to examine the effectiveness of the within-task adaptation approach.

3.1.5. Cross-Task Adaptation Approach

In this phase, QARiB is further pre-trained on a task-specific unlabeled dataset (e.g., sentiment data), and fine-tuned on the labeled dataset from the other task (e.g., emotion dataset) and vice versa. This can be expressed as:

$$QARiB_T \rightarrow PretrainingDataset_T \rightarrow FinetuningDataset_S \quad (2)$$

where S $\neq$ T are the source and target task and can be either sentiment or emotion. The overall process of the cross-task adaptation approach of QARiB is illustrated in Figure 2.

One of the objectives of this experiment is to discover if there is a relationship between sentiment and emotion tasks. To do this, we utilised this approach to determine if the model trained on sentiment data may benefit and enhance the emotion outcomes. In the other direction, we wanted to investigate if the model developed using emotion data may improve and enhance sentiment results. To accomplish this, we used the same pre-trained sentiment models (QST, QSR, and QSRT) as well as emotion models (QE3, and QE6) (described in Section 3.1.4). The sentiment models were fine-tuned using different emotion

datasets, while the emotion models were fine-tuned using various sentiment datasets as illustrated in Figure 2.

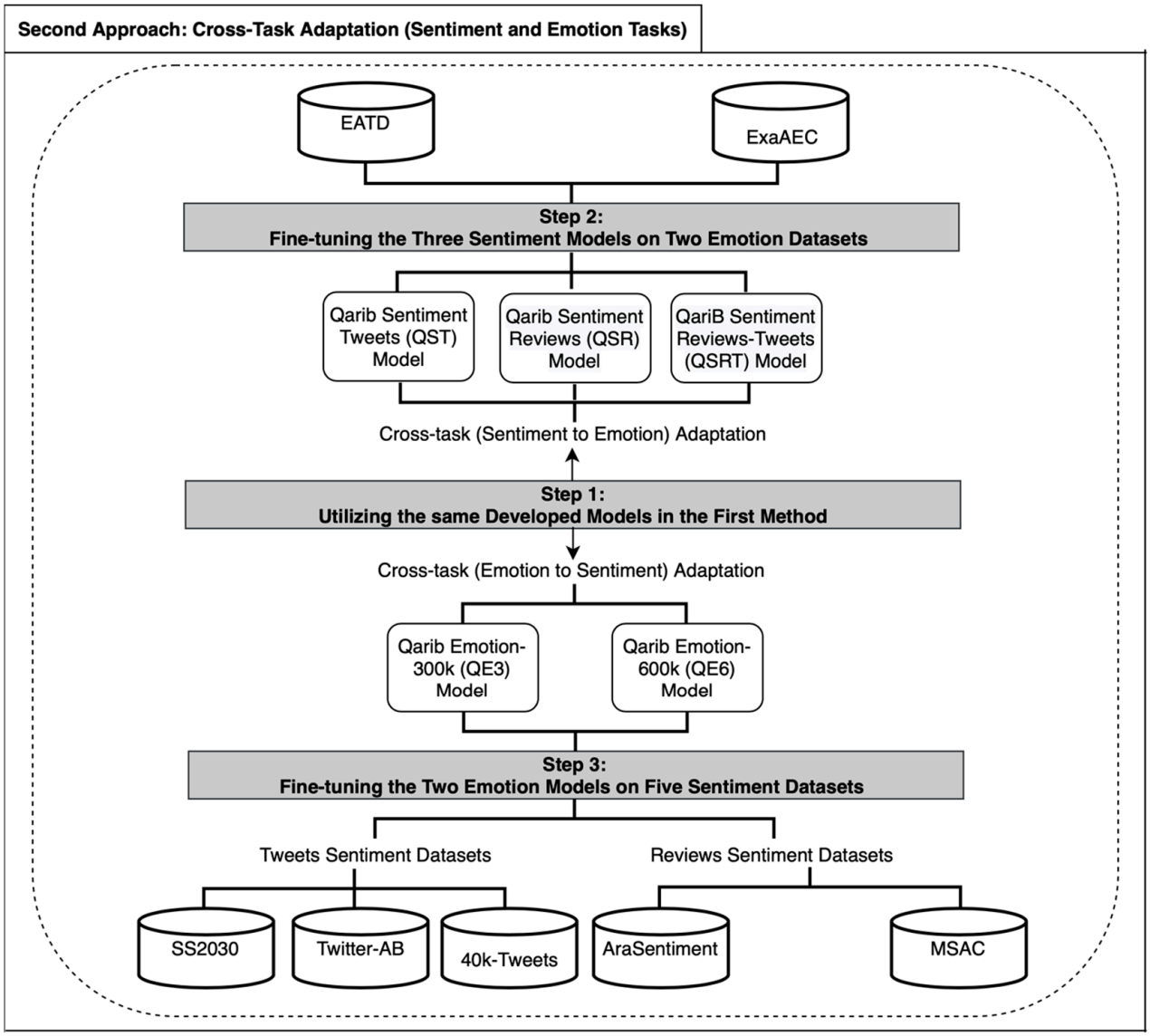

Figure 2. Language Models Development Process Using the Cross-Task Adaptation Approach.

3.2. Performance Evaluation Metrics

In this work, we present the experimental results to ensure the efficiency of our models using the following metrics:

3.2.1. Precision

Precision is defined as the proportion of true positives to all positives.

$$Precision = TP/(TP + FP) \tag{3}$$

3.2.2. Recall

Recall is the proportion of correctly identified examples from a specified class to all examples of that class.

$$Recall = TP/(TP + FN) \tag{4}$$

3.2.3. Accuracy

Accuracy is defined as the proportion of the total number of correct prediction examples to the total number of predictions.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (5)$$

3.2.4. F1-Score

The F1-Score provides one value that combines precision and recall, also known as the harmonic mean. One of the most commonly used metrics, computed as:

$$F1 = 2 \times (precision \times Recall)/(precision + Recall) \quad (6)$$

where:

- Positive examples correctly predicted are denoted by *TP* (true positive)
- Negative examples correctly predicted are denoted by *TN* (true negative)
- Incorrect positive predictions are denoted by *FP* (false positive)
- The wrong negative predictions are denoted by *FN* (false negative).

We report the macro average (also known as the unweighted mean) for each precision, recall, and F1-score. Using the macro average, each class is given equal importance, even if the classes are imbalanced. The results are discussed and compared specifically based on their macro-F1 score.

4. Experiments

We evaluated our adapted models on two Arabic NLP downstream tasks: sentiment analysis and emotion detection. In order to investigate whether further pre-training of the QARiB model [17] using task-specific unlabeled data could continue to improve the performance of the QARiB model [17] on sentiment and emotion tasks, the experiments aimed at addressing the following research questions for this study.

1. Does the type of training data (tweets or reviews) used in QARiB's language model further pre-training stage affect the end-task results?

To provide an answer to RQ1, the sentiment fine-tuning datasets that we used came from two distinct domains (Twitter and reviews). We intended to study the training and fine-tuning using various data types and evaluate model performance on each dataset from different points of view. For illustration, the QST model was trained using just tweets. However, we wanted to study the extent to which it performed well with review datasets (e.g., ArSentiment, and MASC). In contrast, we also wanted to evaluate the performance of the QSR model trained using only reviews on the tweet datasets (e.g., SS2030, 40k-Tweets, Twitter-AB).

2. What sentiment classification performance can be achieved if the QARiB language model is further pre-trained on a sentiment-specific dataset?

To address RQ2, we investigated whether sentiment models such as QST, QSR, and QSRT, which were further pre-trained on unlabeled sentiment datasets, could improve performance when fine-tuned on various labelled sentiment datasets. In this experiment, five sentiment datasets were used to fine-tune sentiment models.

3. What emotion classification performance can be achieved if the QARiB language model is further pre-trained on an emotion-specific dataset?

To find the answer to RQ3, we studied how well emotion models, such as QE3, and QE6 models, that were trained on emotion-unlabeled datasets, performed when fine-tuned on a variety of emotion-labelled datasets. We fine-tuned the emotion models using two-emotion datasets in an attempt to enhance the classification results. Moreover, we fine-tuned the BERT-base-QaRiB model as a baseline model on all seven sentiment and emotion datasets and compared the results.

4. Is there a relationship between sentiment and emotion representation? (i.e., can further pre-training QARiB with a sentiment dataset boost emotion classification results and vice versa?)

To provide an answer to RQ4 and see if there is a relationship between the sentiment and emotion tasks, we fine-tuned sentiment models QST, QSR, and QSRT on the two emotion datasets to examine whether the model trained on sentiment data could improve or increase the performance of emotion classification. Second, we fine-tuned the QE3 and QE6 models on the five sentiment datasets to see whether the model trained using emotion data could improve or enhance the results of the sentiment classification performance.

This section describes the experiment's setup, including the evaluation datasets, the baseline model compared to our models, the fine-tuning architecture, and the hyperparameter choices for fine-tuning our models.

4.1. Fine-Tuning Datasets

The datasets used for the evaluation process were chosen from the available Arabic sentiment and emotion dataset. For fine-tuning our models, we used five sentiment datasets and two emotion datasets. For all fine-tuning experiments, we applied the standard train/development/test set split of 80/10/10. Below is a description of the datasets utilised:

4.1.1. Sentiment Datasets

In order to cover different domains or sources, we chose the five sentiment datasets from different domains, including Twitter and reviews. The SS2030 [75], 40k-Tweets [76], and Twitter-AB [77] datasets were sourced from Twitter. In addition, ArSentiment [78], and MASC [79] were reviews datasets.

- SS2030 dataset [75]: sentiment dataset that has been gathered from Twitter includes 4252 tweets focusing on a variety of social issues in Saudi Arabia. The data set was manually annotated, and it consists of two classes (2436 positive, 1816 negative).
- Twitter-AB [77]: This dataset consists of 2000 tweets that were gathered from Twitter and have been classified into 1k positive, and 1k negative. The dataset was manually labelled and included both MSA and the Jordanian dialect, encompassing diverse topics related to politics and the arts.
- 40k-Tweets [76]: There are 40,000 tweets in this dataset, 20,000 of which are positive and 20,000 of which are negative. These tweets are written in both MSA and an Egyptian dialect. Furthermore, the gathered tweets are manually labelled and span a wide range of topics such as politics, sports, health, and social problems.
- The ArSentiment [78] is a large and multi-domain reviews dataset consisting of over 45k reviews on the 5-rating scale, for movies, hotels, restaurants, and products. We used a rating scale to assign labels to data, 1 and 2 stars have been considered negative, 3 stars have been considered neutral, and 4 and 5 stars have been considered positive.
- Multi-domain Arabic Sentiment Corpus (MASC) [79]: a review dataset that was scraped from a variety of websites including Google Play, Twitter, Facebook, and Qaym. The dataset, which included several different domains, was manually annotated into two classes: positive, and negative.

We selected datasets of varying sizes, some of which contained 40,000 sentences such as 40k-Tweets, and ArSentiment. Others, such as SS2030, Twitter-AB, and MASC had sizes of less than 7000 sentences. The selection of sentiment datasets with diverse domains and sizes was motivated by a desire to examine the impact of adaptation approaches from multiple perspectives. The statistics and classes distribution of the sentiment datasets are shown in Table 6. In addition, Table 7 provides the number of train, development, and test samples for each dataset.

Table 6. Sentiment Datasets Statistics.

Dataset	Size	#Classes	#Positive	#Negative	#Neutral	Source
SS2030	4252	2	2436	1816	-	Twitter
Twitter-AB	1961	2	999	962	-	
40k-Tweets	39,993	2	19,998	19,995	-	
ArSentiment	45,498	3	33,003	9336	3159	Reviews
MASC	6733	2	4476	2257	-	

Table 7. Train/Dev/Test Samples for Sentiment Datasets.

Dataset	Train Samples	Dev Samples	Test Samples
SS2030	3401	426	425
Twitter-AB	1568	197	196
40k-Tweets	31,994	4000	3999
ArSentiment	36,398	4550	4550
MASC	5386	674	673

4.1.2. Emotion Datasets

We evaluated our models on emotion Arabic tweet dataset (EATD) [80] and ExaAEC dataset [81]. In comparison to sentiment datasets, the labelled emotion datasets for the Arabic language are small and scarce. All these datasets are derived from Twitter. Table 8 illustrates the distribution of classes for each dataset. The number of train, development, and test samples for each dataset is presented in Table 9.

Table 8. Emotion Datasets Statistics.

Dataset-Name	EATD	ExaAEC
Size	2021	4738
#Classes	4	6
#Joy	629	472
#Sad	414	1909
#Fear	359	195
#Anger	619	191
#Surprise	-	795
#Disgust	-	1176

Table 9. Train/Dev/Test Samples for Emotion Datasets.

Dataset	Train Samples	Dev Samples	Test Samples
EATD	1616	203	202
ExaAEC	3790	474	474

- EATD [80]: an Arabic emotion dataset gathered from Twitter. The dataset was classified into four classes including anger, disgust, joy, and sadness. The annotation of the dataset was automatically for over 22k tweets based on emojis and manually for a subset of 2021 tweets. The manually annotated dataset has been utilised in our experiments.
- ExaAEC [81]: a multi-label Arabic emotion dataset consisting of approximately 20,000 tweets categorized as "neutral", "joy", "love", "anticipation", "acceptance", "surprise", "sadness", "fear", "anger", and "disgust." Each tweet in this dataset was manually annotated with one or two emotions. Given that the dataset contains tweets with multiple labels, we select a subset containing only tweets with a single label and according to the Ekman model, as follows: 'sadness' 1909, 'disgust' 1176, 'surprise' 795, 'joy' 472, 'fear' 195, 'anger' 191, for a total of approximately 4738 tweets.

4.2. Fine-Tuning Architecture

Fine-tuning the BERT model is "simple and direct", as indicated by [18], and only requires the addition of one more layer after the last BERT layer and training for a small number of iterations. The input sequence used to fine-tune the language model, in this case, is represented by the tokens [CLS] and [SEP] appended to the beginning and end of the sentence, respectively. The [CLS] token is used for all classification-related tasks. As a result, our models can be utilised for a variety of downstream text classification tasks with only minor architecture changes needed. Specifically, we fine-tuned our models for sentiment and emotion classification in Arabic text using the same fine-tuning strategy as BERT [18]. Trainer is a class within the Transformers library that can be utilised to fine-tune a variety of pre-trained Transformers-based models using a specific dataset. For the purpose of instantiating our sequence classification models, we utilised the AutoModelForSequenceClassification class. Due to the fact that our models were not pre-trained on the process of classifying sentences, the head of the model that had been pre-trained was removed, and in its place, a new head more suited to each task was added. The new head's weights were initially selected at random. This indicates that during model fine-tuning, just the weights of the new layers will be updated. In other words, during fine-tuning, all of the layers in our models will be frozen. For classification tasks, we added a fully connected feed-forward layer to the model and used the standard SoftMax activation function for prediction. It is worth noting that we fine-tuned our models independently for each task and dataset, using the same fine-tuning architecture. For a specific number of epochs, we fine-tuned our models on the training set. After that, the model checkpoint with the lowest validation loss was chosen automatically. We then used this checkpoint to do an evaluation of the test set.

4.3. Fine-Tuning Hyper-Parameters

Evaluating or fine-tuning the pre-trained language model is time-consuming, and manually experimenting with various hyperparameters might take days. Hyperparameter optimisation libraries, like Ray-tune [82], allow for the automatic selection of optimal values for model hyperparameters. This library is compatible with a wide variety of machine learning frameworks, including PyTorch and TensorFlow. This library was used in the experiments we conducted for this work. We ran ten trials for each dataset, and the hyperparameters were randomly chosen by the tool. After the hyperparameter search was completed, we obtained the best hyperparameters, which were used to fine-tune our final model. It should be noted that, due to computational and time constraints, we did not run the search for more than ten trials. In fact, for some datasets, such as 40k-Tweets and ArSentiment, the training time ranges from 6 to 10 hours and may exceed that time depending on the hyperparameters chosen by RayTune.

4.4. Baseline Models

As a baseline, to estimate how well our models performed, we compared them to the BERT-base-QaRiB model's [17] performance on the same tasks. We used a currently available BERT-base-QaRiB model and performed supervised fine-tuning, as described in Section 4.2, of the model's parameters for each dataset. Moreover, the results of this study were compared to the benchmark results provided by the datasets' original papers [75–80]. Except for the ExaAEC dataset, of which only a subset was used in this work, which is incompatible with the version used in [81].

5. Results and Discussion

5.1. Exp-I: Experiment to Investigate the Influence of the Within-Task Adaptation Approach on Sentiment and Emotion Classification Performance

The results of fine-tuning the BERT-base-QaRiB, QST, QSR, and QSRT models on the SS2030, Twitter-AB, 40k-Tweets, ArSentiment, and MASC datasets are shown in Table 10. Table 11 presents the results obtained by fine-tuning the BERT-base-QaRiB, QE3, and QE6

models using the EATD and ExaAEC datasets. The results are discussed, compared, and analysed specifically based on the macro-F1 score for the partition of the test set. In addition, the results of each dataset of the base studies [75–80] are presented in Tables 10 and 11. The results that showed an enhancement above the results obtained by the baseline model (i.e., the BERT-base-QaRiB model) are typically highlighted in bold. The results that are highlighted in bold and underlined are the best results that have been achieved for each dataset according to the model used. In total, 26 separate experiments were carried out utilising various sentiment and emotion datasets.

Table 10. Results of fine-tuning sentiment models on sentiment datasets (All-metrics).

Source	Datasets	Models	Precision	Recall	Accuracy	Macro-F1
Twitter	SS2030	**SVM** [75]	90.1%	-	89.83%	89.7%
		BERT-base-QaRiB	90.87%	90.56%	91.06%	90.70%
		QST	92.97%	92.89%	93.18%	**92.93%**
		QSR	91.35%	91.70%	91.76%	**91.51%**
		QSRT	92.05%	92.47%	92.47%	**92.24%**
	Twitter-AB	**SVM** [77]	-	-	87.2%	-
		BERT-base-QaRiB	94.47%	94.32%	94.39%	94.37%
		QST	96.93%	96.93%	96.94%	**96.93%**
		QSR	96.41%	96.49%	96.43%	**96.43%**
		QSRT	97.43%	97.47%	97.45%	**97.45%**
	40k-Tweets	**LSTM** [76]	86.94%	88.9%	88.05%	87.24%
		BERT-base-QaRiB	90.56%	90.51%	90.50%	90.49%
		QST	91.38%	91.37%	91.37%	**91.37%**
		QSR	91.34%	91.27%	91.27%	**91.27%**
		QSRT	90.66%	90.65%	90.65%	**90.65%**
Reviews	ArSentiment	**SVM** [78]	-	-	59.9%	-
		BERT-base-QaRiB	81.21%	72.52%	90.62%	75.93%
		QST	79.04%	75.34%	90.48%	**76.83%**
		QSR	81.61%	76.53%	91.45%	**78.53%**
		QSRT	81.20%	75.83%	91.25%	**78.15%**
	MASC	**LLR** [79]	-	-	-	97.8%
		BERT-base-QaRiB	95.65%	93.81%	95.10%	94.61%
		QST	97.04%	97.20%	97.33%	**97.12%**
		QSR	95.95%	95.37%	95.99%	**95.65%**
		QSRT	96.14%	96.53%	96.58%	**96.33%**

Table 11. Results of fine-tuning emotion models on emotion datasets (All-metrics).

Datasets	Models	Precision	Recall	Accuracy	Macro-F1
EATD	**SVM** [80]	69.67%	69.04%	-	68.52%
	BERT-base-QaRiB	85.31%	86.47%	85.64%	85.46%
	QE3	92.79%	89.63%	90.59%	**90.10%**
	QE6	88.20%	89.12%	88.61%	**88.31%**
ExaAEC	**BERT-base-QaRiB**	65.87%	62.84%	74.26%	63.81%
	QE3	65.53%	63.72%	75.11%	**64.21%**
	QE6	63.80%	64.90%	74.68%	**64.16%**

In Table 10, the results that outperformed the BERT-base-QaRiB model results are highlighted in bold. All sentiment models, including QST, QSR, and QSRT, outperformed the BERT-base-QaRiB model on all sentiment datasets. In addition, when comparing the QST model to the QSR model across all of the experiments (given in Table 10), we observed that the sentiment datasets that were sourced from Twitter, including SS2030, Twitter-AB, and 40k-Tweets, the QST model outperformed the QSR model. Based on this, we may infer

that data distributions of tasks within the same source or domain could be similar. This also indicates that further pre-training of the model using task-specific datasets from the same genre or domain of the fine-tuning datasets yields better results than utilising datasets from a different genre or domain.

Compared to the BERT-base-QaRiB model, Table 10 reveals a performance gain of 2.22% for the QST model and 0.80% for the QSR model on the SS2030 dataset. In addition, the QST and QSR models showed improvements in the Twitter-AB dataset by 2.56% and 2.06%, respectively. In fact, the 40k-Tweets dataset performance increased only by 0.88% using the QST model and by 0.77% using the QSR model. The explanation might be that the 40k-Tweets dataset is a multi-domain dataset including several domains, such as politics and arts, and these domains were not included or covered extensively during the training of the model. Comparing the improvement in the performance of our models QST and QSR on the SS2030 and Twitter-AB datasets to the 40k-Tweets dataset may suggest that the models perform better on small datasets as opposed to large datasets. Compared to the BERT-base-QaRiB model, Table 10 reveals an improvement in the performance of 0.90% for the QST model and 2.60% for the QSR model on the ArSentiment dataset. Meanwhile, performance on the MASC dataset improved by 2.51% using the QST model and by 1.04% using the QSR model.

The QSRT model outperformed BERT-base-QaRiB on the SS2030, Twitter-AB, 40k-Tweets, ArSentiment, and MASC datasets by 1.54%, 3.08, 0.15%, 2.22%, and 1.72%, respectively. Our best sentiment model for SS2030, 40k-Tweets and MASC datasets was the QST model, which achieved 92.93%, 91.37%, and 97.12% F1-scores, respectively. Moreover, the QSR model obtained the highest F1 score on the ArSentiment dataset by achieving 78.53%. The QSRT model was the best sentiment model on the Twitter-AB dataset with macro-F1 of 97.45%. This may suggest that there is no need to perform further pre-training of a model, with a large amount of training data. Instead, training with task-specific datasets that share the same domain as the fine-tuning datasets could result in higher performance.

Table 10 demonstrates that the developed models significantly outperform the results of the original studies for the SS2030, Twitter-AB, 40k-Tweets, and ArSentiment datasets in terms of accuracy by 3.35%, 10.25% 3.32%, and 31.55%, respectively. In terms of the F1 score, it improved by 0.04% on the MASC dataset. The results reported in Table 10 show that the within-task adaptation approach has a beneficial impact on the final results of the sentiment analysis task. In other words, further pre-training of the QARiB language model with unlabeled sentiment datasets and fine-tuning using labelled sentiment datasets improved or enhanced the final results of the sentiment analysis task. In addition, further pre-training using sentiment-specific datasets with the same source or domain as the fine-tuning datasets leads to better enhancement in the sentiment classification results.

In terms of the results of emotion detection, it is challenging for the model that the emotion datasets used, such as EATD and ExaAEC, have multiple classes (i.e., four and six emotion classes). Nevertheless, compared to BERT-base-QaRiB, our QE3 and QE6 models performed better on all emotion datasets, as shown in Table 11 in bold. Compared to the BERT-base-QaRiB model, the QE3 model performed 4.64% better on the EATD dataset. Using the same dataset and the QE6 model, a 2.84% improvement in performance was observed. Results for the ExaAEC dataset were enhanced by 0.40% with the QE3 model and by 0.35% with the QE6 model. The reason could be that the ExaAEC dataset contains six emotion classes in addition to being an unbalanced dataset, as demonstrated in Table 8.

It can be observed that the QE3 model was the best emotion model across all emotion datasets, including the EATD, and ExaAEC datasets, obtaining macro-F1 scores of 90.10%, and 64.21%, respectively. Furthermore, Table 11 shows that the QE3 model outperforms the SVM model in terms of the F1 score by 21.58% for the EATD dataset. These results indicate that QE3 outperformed QE6 on the two emotion datasets. In addition, as shown in Table 5, the training results of the QE6 models were not superior to those of the QE3 models. In fact, QE6 validation loss and perplexity increased. Together, these findings

provide an important insight, namely that pre-training the model for longer training steps is not necessary to achieve optimal performance.

Finally, the results reported in Table 11 show that the within-task adaptation approach has a positive impact on the final results of the emotion detection task. Accordingly, pre-training the QARiB language model with unlabeled emotion datasets and fine-tuning it with labelled emotion datasets improves emotion detection task performance. In addition, further pre-training of the model for longer training steps is unnecessary to get the highest performance.

5.2. Exp-II: Experiment to Investigate the Influence of the Cross-Task Adaptation Approach on Sentiment and Emotion Classification Performance

Table 12 summarises the results of fine-tuning the BERT-base-QaRiB, QE3, and QE6 models using the sentiment datasets SS2030, Twitter-AB, 40k-Tweets, ArSentiment, and MASC. Table 13 shows the results of fine-tuning the BERT-base-QaRiB, QST, QSR, and QSRT models using EATD and ExaAEC datasets. The results are discussed, compared, and analysed specifically based on the macro-F1 score for the partition of the test set. In addition, the results of each dataset of the original studies [75–80] are presented in Tables 12 and 13. The results that showed an improvement over those obtained by the baseline model (i.e., BERT-base-QaRiB model) are highlighted in bold. The results that are in bold and underlined represent the highest results that have been obtained for each dataset and according to which model. In total, 16 experiments were carried out using various sentiment and emotion datasets.

Table 12. Results of fine-tuning emotion models on sentiment datasets (All-metrics).

Source	Datasets	Models	Precision	Recall	Accuracy	Macro-F1
Twitter	SS2030	**SVM** [75]	90.1%	-	89.83%	89.7%
		BERT-base-QaRiB	90.87%	90.56%	91.06%	90.70%
		QE3	93.21%	92.62%	93.18%	**92.89%**
		QE6	91.09%	91.50%	91.53%	**91.27%**
	Twitter-AB	**SVM** [77]	-	-	87.2%	-
		BERT-base-QaRiB	94.47%	94.32%	94.39%	94.37%
		QE3	96.41%	96.49%	96.43%	**96.43%**
		QE6	96.06%	95.83%	95.92%	**95.90%**
	40k-Tweets	**LSTM** [76]	86.94%	88.9%	88.05%	87.24%
		BERT-base-QaRiB	90.56%	90.51%	90.50%	90.49%
		QE3	91.22%	91.19%	91.20%	**91.20%**
		QE6	91.40%	91.40%	91.40%	**91.40%**
Reviews	ArSentiment	**SVM** [78]	-	-	59.9%	-
		BERT-base-QaRiB	81.21%	72.52%	90.62%	75.93%
		QE3	80.57%	77.83%	91.10%	**79.12%**
		QE6	80.25%	77.50%	91.32%	**78.65%**
	MASC	**LLR** [79]	-	-	-	97.8%
		BERT-base-QaRiB	95.65%	93.81%	95.10%	94.61%
		QE3	96.03%	95.95%	96.29%	**95.99%**
		QE6	95.45%	95.60%	95.84%	**95.52%**

Table 13. Results of fine-tuning sentiment models on emotion datasets (all-metrics).

Datasets	Models	Precision	Recall	Accuracy	Macro-F1
EATD	**SVM** [80]	69.67%	69.04%	-	68.52%
	BERT-base-QaRiB	85.31%	86.47%	85.64%	85.46%
	QST	91.58%	89.92%	91.09%	**90.18%**
	QSR	88.18%	88.89%	88.61%	**88.27%**
	QSRT	89.74%	90.26%	90.10%	**89.78%**
ExaAEC	**BERT-base-QaRiB**	65.87%	62.84%	74.26%	63.81%
	QST	67.98%	63.58%	76.79%	**65.05%**
	QSR	67.19%	63.83%	75.53%	**64.98%**
	QSRT	71.33%	64.25%	75.74%	**64.38%**

Overall, as shown in Tables 12 and 13, all sentiment models, including QST, QSR, and QSRT, and emotion models, including QE3, and QE6, outperformed the BERT-base-QaRiB model for all sentiment and emotion datasets. In the tables, results that exceeded the BERT-base-QaRiB model results are highlighted in bold. Table 12 reveals that the QE3 model outperformed BERT-base-QaRiB on the SS2030, Twitter-AB, 40k-Tweets, ArSentiment, and MASC datasets by 2.18%, 2.06%, 0.70%, 3.18%, and 1.38%, respectively. In addition, the QE6 model outperformed BERT-base-QaRiB on the SS2030, Twitter-AB, 40k-Tweets, ArSentiment, and MASC datasets by 0.57%, 1.53%, 0.90%, 2.71%, and 0.91%, respectively.

When comparing emotion models, including QE3 and QE6, across five sentiment datasets, QE3 outperforms QE6 on the SS2030, Twitter-AB, ArSentiment, and MASC datasets, obtaining macro-F1 by 92.89%, 96.43%, 79.12%, and 95.99%, respectively. On the 40k-Tweets dataset, the QE6 model outperforms the QE3 and obtains macro-F1 by 91.40%. These results may provide insight into the fact that pre-training the model for longer training steps does not necessarily give optimal performance. In addition, the further pre-training of the model using sentiment data and fine-tuning it with an emotion dataset can improve the final emotion classification results.

In Table 13, compared to the BERT-base-QaRiB model, the QST model improved performance on the EATD dataset by 4.71%, making it the best model on this dataset with a macro-F1 of 90.18%. On the same dataset and using the QSR model, a 2.80% improvement in performance was observed. On the same dataset, the QSRT model showed a performance improvement of 4.32%, which was better compared to the QSR models. Using the ExaAEC dataset, QST model outperformed all other models with an improvement of 1.24% and a macro-F1 of 65.05%. While using the QSR and QSRT models, performance improvements of 1.17% and 0.57% were achieved.

Comparing the results of the QST and QSR models on emotion datasets, including EATD and ExaAEC on all emotion datasets, we noticed that the QST model outperformed the QSR model. This was expected because the QST model was further trained using tweet data, and the emotion datasets were also taken from Twitter. These findings indicate that we might get better results if we further pre-train the model using a task-specific dataset from the same genre for fine-tuning the datasets. On emotion datasets, the QSRT model performed somewhat worse than the QST and QSR models on the ExaAEC dataset, although the QSRT model was trained with a larger dataset and a different data genre (Twitter and reviews). This may indicate that a large quantity of data may not be necessary for training the model and that a dataset of the same genre as the dataset used for fine-tuning may be more efficient. In general, Table 12 demonstrates that the developed models significantly outperform the results of the base studies for the SS2030, Twitter-AB, 40k-Tweets, and ArSentiment datasets, except the MASC dataset, in terms of accuracy by 3.35%, 9.23%, 3.35%, and 31.42%, respectively. Furthermore, Table 13 shows that the QST model outperforms the SVM model in terms of f1-score by 21.66% for the EATD dataset.

In conclusion, the results presented in Tables 12 and 13 illustrate the effectiveness of the cross-task adaptation approach on the final results of the sentiment and emotion classification tasks. These results suggest that the data distribution between sentiment and emotion may be converging. We can see how each task can influence and improve the results of the other. This may give an important insight into how convergent tasks with converged data distribution might enhance each other's performance. For instance, the Arabic emotion detection task has more limited resources than sentiment. Therefore, this study may help researchers tackling emotion detection to obtain better results by utilising sentiment resources. Additionally, when comparing the results of the two task-adaptation approaches (cross-task and within-task), it can be shown that the cross-task adaptation results sometimes outperform the within-task approach. On the emotion datasets EATD and ExaAEC, for instance, the sentiment model QST outperformed the emotion models QE3 and QE6. Additionally, the emotion models QE3 and QE6 outperformed the sentiment models on the sentiment datasets 40k-Tweets and ArSentiment.

6. Conclusions

The experiments described in the previous sections examine the effect of two adaptation approaches: within-task and cross-task adaptation. In total, five new models were developed using the previous approach: the QST, QSR, QSRT, QE3, and QE6 models. Different evaluation experiments were conducted by fine-tuning each model for two downstream tasks, sentiment analysis and emotion detection. Using five sentiment datasets, including SS2030, Twitter-AB, 40k-Tweets, ArSentiment, and MASC, in addition to two emotion datasets, EATD and ExaAEC, 42 experiments were carried out in total. The sentiment and emotion datasets covered both small- and large-resource settings. The experiments reveal the following: first, the within-task and cross-task adaptation approaches have influenced the final results and boosted performance for all tasks (i.e., sentiment and emotion). Second, our newly developed QST, QSR, QSRT, QE3, and QE6 models outperformed the BERT-base-QaRiB model on all sentiment and emotion datasets. Third, the training using task-specific datasets that share the same domain as the fine-tuning datasets results in higher performance. Fourth, additional pre-training of the model for longer training steps is unnecessary to get the highest performance. Finally, cross-task adaptation shows that sentiment and emotion data may converge, and each task might enhance the results of the other.

This study showed that pre-training the QARiB language model on small-scale sentiment or emotion data improves model understanding of this domain data and yields considerable improvements. Because of the scarcity of emotion datasets, one of the limitations of this research is that the model was only evaluated on two small emotion datasets. In general, findings reveal interesting areas for future research. The findings indicate that these approaches (i.e., within-task and cross-task adaptation) can improve the performance of QARiB. Consequently, any pre-trained Arabic language model can be utilised with the approaches that we have investigated. While Arabic language models like AraBERT and MARBERT already perform effectively well on sentiment and emotion tasks, they may benefit significantly from further task-specific pre-training. In addition, we believe that pre-training on larger task-specific data could further enhance performance. Finally, the developed language models are publicly available to be used by the NLP community for research purposes, and we hope this work helps researchers interested in the domain of Arabic sentiment and emotion analysis.

Author Contributions: Conceptualisation, W.A., N.A.-T. and A.A.; data curation, W.A.; formal analysis, W.A.; funding acquisition, A.A.; methodology, W.A., N.A.-T. and A.A.; software, W.A.; supervision, N.A.-T. and A.A.; validation, W.A.; writing—original draft, W.A.; writing—review and editing, N.A.-T. and A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Deanship of Scientific Research, King Saud University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets have been taken from this link https://arbml.github.io/masader/ (accessed on 7 September 2022).

Acknowledgments: The authors are grateful to the Deanship of Scientific Research, King Saud University for funding through the Vice Deanship of Scientific Research Chairs.

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

References

1. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
2. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
3. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]

4. Torrey, L.; Shavlik, J. Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*; IGI Global: Hershey, PA, USA, 2010; pp. 242–264.
5. Oueslati, O.; Cambria, E.; Ben HajHmida, M.; Ounelli, H. A review of sentiment analysis research in Arabic language. *Future Gener. Comput. Syst.* **2020**, *112*, 408–430. [CrossRef]
6. Abdullah, M.; Hadzikadicy, M.; Shaikhz, S. SEDAT: Sentiment and emotion detection in Arabic text using CNN-LSTM deep learning. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 17–20 December 2018; pp. 835–840.
7. Stevenson, A. *Oxford Dictionary of English*; Oxford University Press: New York, NY, USA, 2010.
8. Ekman, P.; Friesen, W.V.; O'sullivan, M.; Chan, A.; Diacoyanni-Tarlatzis, I.; Heider, K.; Krause, R.; LeCompte, W.A.; Pitcairn, T.; Ricci-Bitti, P.E. Universals and cultural differences in the judgments of facial expressions of emotion. *J. Pers. Soc. Psychol.* **1987**, *53*, 712. [CrossRef]
9. Plutchik, R. A general psychoevolutionary theory of emotion. In *Theories of Emotion*; Elsevier: Amsterdam, The Netherlands, 1980; pp. 3–33.
10. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef] [PubMed]
11. Alsentzer, E.; Murphy, J.; Boag, W.; Weng, W.-H.; Jindi, D.; Naumann, T.; McDermott, M. Publicly Available Clinical BERT Embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, MN, USA, 5 June 2019; pp. 72–78.
12. Araci, D. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv* **2019**, arXiv:1908.10063.
13. Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; Smith, N.A. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020.
14. Sun, C.; Qiu, X.; Xu, Y.; Huang, X. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 194–206.
15. Ameur, M.S.H.; Aliane, H. AraCOVID19-MFH: Arabic COVID-19 Multi-label Fake News & Hate Speech Detection Dataset. *Procedia Comput. Sci.* **2021**, *189*, 232–241.
16. Lan, W.; Chen, Y.; Xu, W.; Ritter, A. An Empirical Study of Pre-trained Transformers for Arabic Information Extraction. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 19–20 November 2020; pp. 4727–4734.
17. Abdelali, A.; Hassan, S.; Mubarak, H.; Darwish, K.; Samih, Y. Pre-training bert on arabic tweets: Practical considerations. *arXiv* **2021**, arXiv:2102.10684.
18. 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, 2–7 June 2019; pp. 4171–4186.
19. Dadas, S.; Perełkiewicz, M.; Poświata, R. Pre-training polish transformer-based language models at scale. In Proceedings of the Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, 12–14 October 2020; Part II 19. Springer: Berlin/Heidelberg, Germany, 2020; pp. 301–314.
20. Polignano, M.; Basile, P.; de Gemmis, M.; Semeraro, G.; Basile, V. AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In Proceedings of the CLiC-it, Bari, Italy, 13–15 November 2019.
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, .; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
22. Antoun, W.; Baly, F.; Hajj, H. AraBERT: Transformer-based Model for Arabic Language Understanding. In Proceedings of the LREC 2020 Workshop Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; p. 9.
23. ElJundi, O.; Antoun, W.; El Droubi, N.; Hajj, H.; El-Hajj, W.; Shaban, K. hulmona: The universal language model in arabic. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, Florence, Italy, 1 August 2019; pp. 68–77.
24. 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.
25. Wadhawan, A. AraBERT and Farasa Segmentation Based Approach For Sarcasm and Sentiment Detection in Arabic Tweets. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, Kyiv, Ukraine, 19 April 2021; pp. 395–400.
26. Antoun, W.; Baly, F.; Hajj, H. AraELECTRA: Pre-Training Text Discriminators for Arabic Language Understanding. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, Kyiv, Ukraine, 19 April 2021; pp. 191–195.
27. Abdul-Mageed, M.; Elmadany, A. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 1–6 August 2021; pp. 7088–7105.
28. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, É.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 8440–8451.

29. Inoue, G.; Alhafni, B.; Baimukan, N.; Bouamor, H.; Habash, N. The interplay of variant, size, and task type in Arabic pre-trained language models. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, Kyiv, Ukraine, 19 April 2021; pp. 32–104.
30. Abdaoui, A.; Berrimi, M.; Oussalah, M.; Moussaoui, A. Dziribert: A pre-trained language model for the algerian dialect. *arXiv* **2021**, arXiv:2109.12346.
31. Alduailej, A.; Alothaim, A. AraXLNet: Pre-trained language model for sentiment analysis of Arabic. *J. Big Data* **2022**, *9*, 1–21. [CrossRef]
32. 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.
33. Elmadany, A.; Nagoudi, E.M.B.; Abdul-Mageed, M. ORCA: A Challenging Benchmark for Arabic Language Understanding. *arXiv* **2022**, arXiv:2212.10758.
34. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
35. Khaddaj, A.; Hajj, H.; El-Hajj, W. Improved generalization of arabic text classifiers. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, Florence, Italy, 1 August 2019; pp. 167–174.
36. El Mekki, A.; El Mahdaouy, A.; Berrada, I.; Khoumsi, A. Domain adaptation for Arabic cross-domain and cross-dialect sentiment analysis from contextualized word embedding. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; pp. 2824–2837.
37. Alqahtani, Y.; Al-Twairesh, N.; Alsanad, A. A Comparative Study of Effective Domain Adaptation Approaches for Arabic Sentiment Classification. *Appl. Sci.* **2023**, *13*, 1387. [CrossRef]
38. Badaro, G.; Jundi, H.; Hajj, H.; El-Hajj, W.; Habash, N. Arsel: A large scale arabic sentiment and emotion lexicon. In Proceedings of the OSACT 3: The 3rd Workshop on Open-Source Arabic Corpora and Processing Tools, Miyazaki, Japan, 8 May 2018; p. 26.
39. Mohammad, S.; Bravo-Marquez, F.; Salameh, M.; Kiritchenko, S. Semeval-2018 task 1: Affect in tweets. In Proceedings of the 12th International Workshop on Semantic Evaluation, New Orleans, LA, USA, 5–6 June 2018; pp. 1–17.
40. Badaro, G.; El Jundi, O.; Khaddaj, A.; Maarouf, A.; Kain, R.; Hajj, H.; El-Hajj, W. Ema at semeval-2018 task 1: Emotion mining for arabic. In Proceedings of the 12th International Workshop on Semantic Evaluation, New Orleans, LA, USA, 5–6 June 2018; pp. 236–244.
41. Aljwari, F. Emotion Detection in Arabic Text Using Machine Learning Methods. *IJISCS-Int. J. Inf. Syst. Comput. Sci.* **2022**, *6*, 175–185.
42. Khalil, E.A.H.; El Houby, E.M.F.; Mohamed, H.K. Deep learning for emotion analysis in Arabic tweets. *J. Big Data* **2021**, *8*, 1–15. [CrossRef]
43. Abdul-Mageed, M.; Zhang, C.; Hashemi, A. AraNet: A Deep Learning Toolkit for Arabic Social Media. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, Marseille, France, 12 May 2020; pp. 16–23.
44. Elfaik, H. Combining context-aware embeddings and an attentional deep learning model for Arabic affect analysis on twitter. *IEEE Access* **2021**, *9*, 111214–111230. [CrossRef]
45. Mansy, A.; Rady, S.; Gharib, T. An ensemble deep learning approach for emotion detection in arabic tweets. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 01304112. [CrossRef]
46. Al-Twairesh, N. The evolution of language models applied to emotion analysis of Arabic tweets. *Information* **2021**, *12*, 84. [CrossRef]
47. Soliman, A.B.; Eissa, K.; El-Beltagy, S.R. Aravec: A set of arabic word embedding models for use in arabic nlp. *Procedia Comput. Sci.* **2017**, *117*, 256–265. [CrossRef]
48. Talafha, B.; Ali, M.; Za'ter, M.E.; Seelawi, H.; Tuffaha, I.; Samir, M.; Farhan, W.; Al-Natsheh, H. Multi-dialect Arabic BERT for Country-level Dialect Identification. In Proceedings of the Fifth Arabic Natural Language Processing Workshop, Barcelona, Spain, 12 December 2020; pp. 111–118.
49. Mahmoud, A.E.-S.; Lazem, S.; Abougabal, M. *Benchmarking a Large Twitter Dataset for Arabic Emotion Analysis*; Research Square: Durham, NC, USA, 2022.
50. Abdelali, A.; Darwish, K.; Durrani, N.; Mubarak, H. Farasa: A fast and furious segmenter for arabic. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, San Diego, CA, USA, 12–17 June 2016; pp. 11–16.
51. Parker, R.; Graff, D.; Chen, K.; Kong, J.; Maeda, K. "Arabic Gigaword." LDC Catalog No. LDC2009T30. 2009. Available online: https://catalog.ldc.upenn.edu/LDC2009T30 (accessed on 1 March 2022).
52. El-Khair, I.A. 1.5 billion words arabic corpus. *arXiv* **2016**, arXiv:1611.04033.
53. Lison, P.; Tiedemann, J. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, 23–28 May 2016.
54. Sennrich, R.; Haddow, B.; Birch, A. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; pp. 1715–1725.

55. Alomari, K.M.; ElSherif, H.M.; Shaalan, K. Arabic tweets sentimental analysis using machine learning. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 602–610.
56. Elmadany, A.; Mubarak, H.; Magdy, W. Arsas: An arabic speech-act and sentiment corpus of tweets. *OSACT* **2018**, *3*, 20.
57. Baly, R.; Khaddaj, A.; Hajj, H.; El-Hajj, W.; Shaban, K.B. ArSentD-LEV: A Multi-Topic Corpus for Target-based Sentiment Analysis in Arabic Levantine Tweets. In Proceedings of the OSACT 3: The 3rd Workshop on Open-Source Arabic Corpora and Processing Tools, Miyazaki, Japan, 8 May 2018; p. 37.
58. Boujou, E.; Chataoui, H.; El Mekki, A.; Benjelloun, S.; Chairi, I.; Berrada, I. An open access NLP dataset for Arabic dialects: Data collection, labeling, and model construction. *arXiv* **2021**, arXiv:2102.11000.
59. Salameh, M.; Mohammad, S.; Kiritchenko, S. Sentiment after translation: A case-study on arabic social media posts. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, CO, USA, 31 May–5 June 2015; pp. 767–777.
60. Kwaik, K.A.; Chatzikyriakidis, S.; Dobnik, S.; Saad, M.; Johansson, R. An Arabic tweets sentiment analysis dataset (ATSAD) using distant supervision and self training. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, Marseille, France, 12 May 2020; pp. 1–8.
61. Nabil, M.; Aly, M.; Atiya, A. Astd: Arabic sentiment tweets dataset. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 2515–2519.
62. Rosenthal, S.; Farra, N.; Nakov, P. SemEval-2017 task 4: Sentiment analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 502–518.
63. Farha, I.A.; Magdy, W. From arabic sentiment analysis to sarcasm detection: The arsarcasm dataset. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, Marseille, France, 12 May 2020; pp. 32–39.
64. Abdellaoui, H.; Zrigui, M. Using tweets and emojis to build tead: An Arabic dataset for sentiment analysis. *Comput. Sist.* **2018**, *22*, 777–786. [CrossRef]
65. Alowisheq, A.; Al-Twairesh, N.; Altuwaijri, M.; Almoammar, A.; Alsuwailem, A.; Albuhairi, T.; Alahaideb, W.; Alhumoud, S. MARSA: Multi-domain Arabic resources for sentiment analysis. *IEEE Access* **2021**, *9*, 142718–142728. [CrossRef]
66. Al-Twairesh, N.; Al-Khalifa, H.; Al-Salman, A.; Al-Ohali, Y. Arasenti-tweet: A corpus for arabic sentiment analysis of saudi tweets. *Procedia Comput. Sci.* **2017**, *117*, 63–72. [CrossRef]
67. Al Omari, M.; Al-Hajj, M.; Hammami, N.; Sabra, A. Sentiment classifier: Logistic regression for arabic services' reviews in lebanon. In Proceedings of the 2019 International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 3–4 April 2019; pp. 1–5.
68. Aly, M.; Atiya, A. Labr: A large scale arabic book reviews dataset. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Sofia, Bulgaria, 22–27 May 2013; pp. 494–498.
69. Elnagar, A.; Khalifa, Y.S.; Einea, A. Hotel Arabic-reviews dataset construction for sentiment analysis applications. In *Intelligent Natural Language Processing: Trends and Applications*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 35–52.
70. Elnagar, A.; Lulu, L.; Einea, O. An annotated huge dataset for standard and colloquial arabic reviews for subjective sentiment analysis. *Procedia Comput. Sci.* **2018**, *142*, 182–189. [CrossRef]
71. Al-Khatib, A.; El-Beltagy, S.R. Emotional tone detection in arabic tweets. In *International Conference on Computational Linguistics and Intelligent Text Processing*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 105–114.
72. Alhuzali, H.; Abdul-Mageed, M.; Ungar, L. Enabling Deep Learning of Emotion With First-Person Seed Expressions. In Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, New Orleans, LA, USA, 6 June 2018; pp. 25–35. [CrossRef]
73. Alqahtani, G. Multimodal Emotion Detection of Social Networks Data Using Deep Learning. Master's Thesis, University of Patras, Patras, Greece, 2022.
74. Beltagy, I.; Lo, K.; Cohan, A. SciBERT: A pretrained language model for scientific text. *arXiv* **2019**, arXiv:1903.10676.
75. Alyami, S.N.; Olatunji, S.O. Application of Support Vector Machine for Arabic Sentiment Classification Using Twitter-Based Dataset. *J. Inf. Knowl. Manag.* **2020**, *19*, 1–13. [CrossRef]
76. Mohammed, A.; Kora, R. Deep learning approaches for Arabic sentiment analysis. *Soc. Netw. Anal. Min.* **2019**, *9*, 52. [CrossRef]
77. Abdulla, N.A.; Ahmed, N.A.; Shehab, M.A.; Al-Ayyoub, M. Arabic sentiment analysis: Lexicon-based and corpus-based. In Proceedings of the 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Amman, Jordan, 3–5 December 2013; pp. 1–6.
78. ElSahar, H.; El-Beltagy, S.R. Building large arabic multi-domain resources for sentiment analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 23–34.
79. Al-Moslmi, T.; Albared, M.; Al-Shabi, A.; Omar, N.; Abdullah, S. Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis. *J. Inf. Sci.* **2018**, *44*, 345–362. [CrossRef]
80. Hussien, W.A.; Tashtoush, Y.M.; Al-Ayyoub, M.; Al-Kabi, M.N. Are emoticons good enough to train emotion classifiers of arabic tweets? In Proceedings of the 2016 7th International Conference on Computer Science and Information Technology (CSIT), Amman, Jordan, 13–14 July 2016; pp. 1–6.

81. Sarbazi-Azad, S.; Akbari, A.; Khazeni, M. ExaAEC: A New Multi-label Emotion Classification Corpus in Arabic Tweets. In Proceedings of the 2021 11th International Conference on Computer Engineering and Knowledge (ICCKE), Mashhad, Iran, 28–29 October 2021; pp. 465–470.
82. Liaw, R.; Liang, E.; Nishihara, R.; Moritz, P.; Gonzalez, J.E.; Stoica, I. Tune: A research platform for distributed model selection and training. *arXiv* **2018**, arXiv:1807.05118.