applied sciences

MDPI

Article

Location Analysis for Arabic COVID-19 Twitter Data Using Enhanced Dialect Identification Models

Nader Essam [1], Abdullah M. Moussa [2,*], Khaled M. Elsayed [2], Sherif Abdou [2], Mohsen Rashwan [3], Shaheen Khatoon [4], Md. Maruf Hasan [4], Amna Asif [4] and Majed A. Alshamari [4]

[1] The Engineering Company for the Development of Digital Systems, Giza 12311, Egypt; naderessam110@gmail.com
[2] Faculty of Computers and Artificial Intelligence, Cairo University, Giza 12613, Egypt; k.mostafa@fci-cu.edu.eg (K.M.E.); s.abdou@fci-cu.edu.eg (S.A.)
[3] Faculty of Engineering, Cairo University, Giza 12613, Egypt; mrashwan@ieee.org
[4] College of Computer Sciences and Information Technology, King Faisal University, AlAhsa 31982, Saudi Arabia; ssyed@kfu.edu.sa (S.K.); mhasan@kfu.edu.sa (M.M.H.); aarkhan@kfu.edu.sa (A.A.); smajed@kfu.edu.sa (M.A.A.)
* Correspondence: a.m.moussa@ieee.org

Abstract: The recent surge of social media networks has provided a channel to gather and publish vital medical and health information. The focal role of these networks has become more prominent in periods of crisis, such as the recent pandemic of COVID-19. These social networks have been the leading platform for broadcasting health news updates, precaution instructions, and governmental procedures. They also provide an effective means for gathering public opinion and tracking breaking events and stories. To achieve location-based analysis for social media input, the location information of the users must be captured. Most of the time, this information is either missing or hidden. For some languages, such as Arabic, the users' location can be predicted from their dialects. The Arabic language has many local dialects for most Arab countries. Natural Language Processing (NLP) techniques have provided several approaches for dialect identification. The recent advanced language models using contextual-based word representations in the continuous domain, such as BERT models, have provided significant improvement for many NLP applications. In this work, we present our efforts to use BERT-based models to improve the dialect identification of Arabic text. We show the results of the developed models to recognize the source of the Arabic country, or the Arabic region, from Twitter data. Our results show 3.4% absolute enhancement in dialect identification accuracy on the regional level over the state-of-the-art result. When we excluded the Modern Standard Arabic (MSA) set, which is formal Arabic language, we achieved 3% absolute gain in accuracy between the three major Arabic dialects over the state-of-the-art level. Finally, we applied the developed models on a recently collected resource for COVID-19 Arabic tweets to recognize the source country from the users' tweets. We achieved a weighted average accuracy of 97.36%, which proposes a tool to be used by policymakers to support country-level disaster-related activities.

Keywords: BERT models; dialect identification; location analysis; language identification; social networks

Citation: Essam, N.; Moussa, A.M.; Elsayed, K.M.; Abdou, S.; Rashwan, M.; Khatoon, S.; Hasan, M.M.; Asif, A.; Alshamari, M.A. Location Analysis for Arabic COVID-19 Twitter Data Using Enhanced Dialect Identification Models. *Appl. Sci.* **2021**, *11*, 11328. https://doi.org/10.3390/app112311328

Academic Editors: Anton Civit and Manuel Dominguez-Morales

Received: 8 October 2021
Accepted: 23 November 2021
Published: 30 November 2021

1. Introduction

On 30 January 2020, the World Health Organization declared COVID-19 a pandemic after the massive spread of the virus SARS-CoV-2 in many countries all over the world [1]. Up to the time of this study, the number of confirmed cases is over 252 million [2]. Such a problem has encouraged many governments to initiate plans of crisis management. One of the most essential rules to reduce the spread of the virus is social distancing. This process encourages individuals to rely more on social network communications. Twitter is arguably one of the most famous social media sites, thanks to its simplicity

and ease of knowledge sharing. The social interactions represent a valuable resource for monitoring and analyzing the impacts of disasters such as COVID-19. Locations of widespread infections, people's reactions to quarantine procedures, symptoms of the new variants of the virus, and the psychological effects can be highlighted from the text analysis of social media data to support real-time disaster operation and management [3–6]. In [7], the author states that the ministry of health of Saudi Arabia has used several accounts on Twitter to populate many health-related hashtags to provide governmental pieces of advice to Saudi Arabian citizens. This is evidence that decision-makers are considering social media as important channels for communicating with people. So, any accurate tool that can provide the policymakers with the response of their specific citizens would be of special importance. And this information can be extracted from the dialectical analysis of social media texts. Furthermore, the identification of Arabic dialects is considered as one of the first pre-processing components in any reliable Arabic Natural Language Processing (NLP) application [8]. This identification is significant for several NLP problems such as information retrieval, automatic translation, sentiment analysis, and event detection. For example, when new cases of spreading disease start to appear in a specific region, the information about such cases most probably begins to emerge on social media. Detecting such events as soon as possible can be very critical for the containment of the disease. And this vital process can be handled using a robust NLP solution.

In this work, we introduce our efforts for analyzing COVID-19 Arabic conversations on the Twitter network. We used a large dataset of Arabic tweets related to the topic of COVID-19 [7]. It has been crawled in the period from January to April 2020. The total number of tweets in the dataset is around 3.8 million. The dataset consists of different Arabic dialects such as Egyptian, Gulf, Levantine, along with Modern Standard Arabic (MSA). The dataset is primarily from hashtags populated from Saudi Arabia, which can be utilized in other Arabic-speaking countries. The dataset includes several types of conversations discussing different aspects of COVID-19 such as social solidarity and preventive measures announced by governments.

It is worth noting that a word in the Arabic language may refer to a wide collection of meanings in different dialects. Officials usually use MSA, the formal version of the language, in educational organizations and pan-Arab news broadcasting, which is different from the varieties that are spoken in daily communications by native speakers [9–11]. These daily varieties constitute the dialects of Arabic that can be classified based on some common linguistic features of geographical locations. The most common Arabic dialects categorization is:

- Egyptian (EGY): An Egyptian dialect spoken in Egypt and is also commonly spoken because of the historical influence of Egyptian media.
- Gulf (GLF): A dialect spoken mostly in Saudi Arabia, the United Arab Emirates, Kuwait, Oman, Bahrain, and Qatar.
- Iraqi: The widely-spoken Arabic dialect by the people of Iraq.
- Levantine (LEV): A dialect spoken mainly by people from the Levant (in other words, people of Palestine, Jordan, Lebanon, and Syria).
- Maghrebi: A dialect of Arabic spoken in North Africa, except Egypt.

Figure 1 illustrates the geographical layout of the above Arabic dialects [12]. We implemented an Arabic dialect identification system to localize COVID-19 data so as to provide region-based insights and analysis. Dialect identification is a form of language identification in which the purpose is to differentiate between dialects that are closely related. Our experiments provided 3.4% absolute enhancement over the state-of-the-art result [13] in dialect identification accuracy on the region level. Additionally, we achieved more than 3% accuracy improvement between the three major Arabic dialects after excluding the MSA set. Moreover, the developed models have been successfully applied on a recently collected dataset of COVID-19-related Arabic tweets to identify the source country of each tweet with a weighted average accuracy of 97.36%. We believe that the achieved accuracy is notable and can be effectively used to support decision-makers

with a valuable source of information regarding people's reactions towards the current pandemic. The rest of the paper is organized as follows: Section 2 reviews the previous effort for Arabic Dialect identification, Section 3 describes the implemented models, and Section 4 includes the description of the used data sets and the results. Finally, Section 5 concludes the paper with future directions.

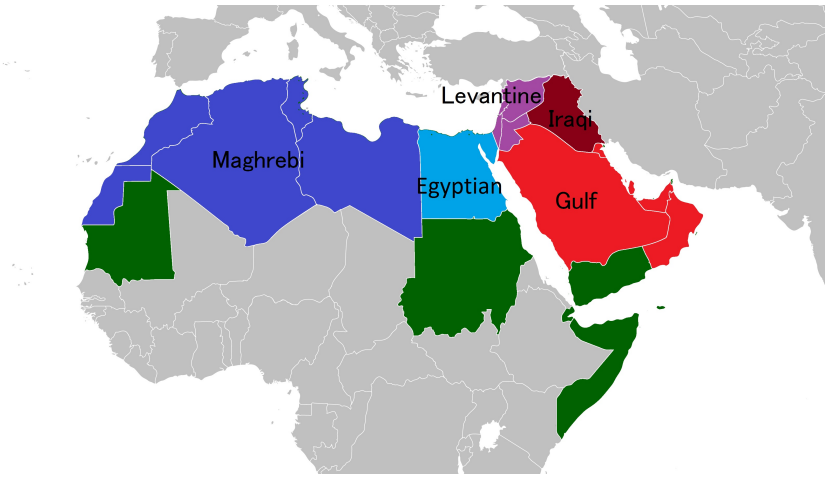

Figure 1. The map of Arabic dialect groups distribution.

2. Related Work

The identification of Arabic dialects is based on both spoken [14–16] and written form [12,13,17–21]. Much of early research has focused on distinguishing between MSA and EGY forms. For example, in [17], Elfardy and Diab suggested a supervised method for the task of sentence-level MSA-EGY identification by exploiting a subset of the Arabic Online Commentary (AOC) dataset. They studied whether pre-processing affects the performance of the classifier or not. The authors found that this was helpful in some cases. They reported an accuracy of 85.5% using 10-fold cross-validation with an SVM classifier. In [12], Zaidan and Callison-Burch tested a word trigram model for MSA, EGY, GLF, and LEV sentences. They achieved an accuracy of 77.8% to classify tweets between EGY and MSA. In the case of (EGY vs. GLF vs. LEV), an accuracy of 83.3% was achieved. The accuracy in the case of the 4-way classification task (MSA, EGY, GLF, and LEV) was 69.4% [22].

In [23], Elfardy et al. suggested a system that identifies linguistic code-switching for the MSA-EGY classification task. The authors adjusted their system in [18] and used a word-based 5-gram model. They utilized a morphological analyzer to annotate the dataset. For annotation, they used MADAMIRA that has been proposed in [24]. The authors improved system in [18] achieved an accuracy of 87.7% when distinguishing between EGY and MSA. In [21], Darwish et al. proposed a work that is similar to these systems in focusing on the classification problem of MSA-EGY for Twitter data. They collected a dataset of 880 K tweets to train the model and used 700 tweets for testing. The authors investigated several morphological and lexical features. They reported a 10% absolute gain over the models trained solely on n-grams.

In [25], Ragab et al. proposed an ensemble model consisting of a group of several classifiers using a set of features that includes word-level and character-level Term Frequency Inverse Document Frequency (TF-IDF) features, class probabilities of many linear classifiers, and probabilities of a language model [26]. Also, Ghoul and Lejeune in [27] presented MICHAEL, a lightweight approach developed for domain dialect identification. Their proposed method used character-level features and perform classification without any pre-processing. They used character N-grams extracted from the original sentences to train a multinomial naive Bayes classifier. MICHAEL achieved an accuracy of 53.25% with N equals to 1, 2, or 3 and accuracy of 62.17% with character 4-grams. In [28], Priban and

Taylor proposed systems that are based on language modeling. They used language models to extract features that are later utilized as an input for other machine learning methods. In [29] Harrat et al. used a classification method based on symmetric Kullback-Leibler For discriminating between dialects in a multi-dialect context.

Huang [30] concentrated on the AOC categories for the 4-way classification task (MSA, EGY, GLF, and LEV). The author claimed that a simple word-level n-gram model trained on the manually annotated portion of AOC as well as unannotated Facebook data improved classification accuracy. On 10% of the manually annotated AOC dataset, Huang [30] reported an accuracy of 87.8%. In [13], Elaraby and Abdul-Mageed proposed a model that handles the binary (MSA, non-MSA), 3-way (EGY, GLF, and LEV), and 4-way (MSA, EGY, GLF, and LEV) classification tasks. They also used the AOC data and reported an accuracy of 87.23%, 87.41%, and 82.45% for binary, 3-way, and 4-way classification tasks, respectively. It is worth mentioning that Huang's [30] result is based on AOC data plus other unannotated data. However, the segmentation of data and benchmark results on AOC is not available. While using AOC in our work, the results of Huang's [30] works are not comparable to our results.

3. Arabic Dialect Identification Model

The proposed dialect identification model consists of two components. The first one is the feature extractor and the second is the classifier system. We looked at several alternatives for each of these two elements in the following two sections.

3.1. The Feature Extractor

The feature extractor module matches input text to a set of features to provide better discriminant representation for the target classes in the mapped dimension space. We investigated two types of features; the first one is based on the TFIDF features [31]. TFIDF measures the relevancy of a word to a specific dialect represented by a group of documents for each dialect. Usually, the regional dialects are identified with compound expressions that consist of more than one word. So, for our feature set, we investigated using the unigram, bigrams, and trigrams expressions. We also investigated using different sizes of relevancy lists ranging from the most important 1000 keywords up to the most important 20,000 words.

The second sort of feature relies on word embedding [32]. Word embedding maps the identities of the words from the discrete domain to the continuous domain, represented by word vectors that can be used as a lexical and semantic representation for words. Different embedding models were investigated. Such models have been developed using different datasets with different characteristics and are described in Section 4.

3.2. The Classifier System

For dialect identification application, we investigated several types of classifiers. We started with the classical logistic regression model. Then we investigated the more advanced neural network-based models. The following subsections provide more details on each one of these classifiers.

3.2.1. Logistic Regression (LR)

Logistic regression is one of the classification algorithms. It can be used to estimate output based on a number of independent parameters representing the used features set. For the LR classifier, we tested the TFIDF features.

3.2.2. LSTM and BiLSTM

The Recurrent Neural Networks (RNN) is one type of effective model for sequential data classification, such as our application of dialect identification of a stream of text. For long sequences, such as the text input on social networks, RNN has the drawback of vanishing gradients. Long Short Term Memory (LSTM) neural networks are designed

to extract long-term dependencies by integrating a memory state into a regular RNN. The LSTM uses a previous state h_(t−1) and an input x_t to measure the hidden state h_t, as illustrated in Figure 2. We used a word-based LSTM in this study, with an architecture of 100 hidden units in 100 dimensions.

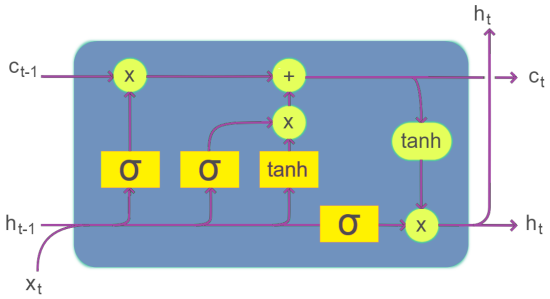

Figure 2. An example of Long Short Term Memory (LSTM) neural network.

Conventional LSTMs have the drawback of being able to decide predictions based only on previously seen content. Bidirectional LSTMs (BiLSTMs) are a form of LSTM that solves this problem by data processing in both directions in two different hidden layers. These two hidden layers are feedforwarded to the same output layer. The forward hidden sequence, the backward hidden sequence, and the output sequence are all computed by BILSTMs.

The architecture we used for the BiLSTM classifier is:

- Embedding Layer: Each word represented with 300 dimensions;
- Sequence length 30 time stamp;
- Dropout rate = 0.8;
- L2 norm of regularization;
- Relu activation;
- Dense Layer: 4 units with a softmax activation function.

In addition, with the recent notable success of combining attention mechanisms with neural networks in a variety of NLP problems, including speech recognition, machine translation, and image captioning [33,34], we investigated employing an attention mechanism with our BiLSTM models to boost the models' accuracy. We used an implementation inspired by [35], in which we applied attention to the LSTM layers output vector.

The architecture we used was the following: Tanh activation function on the outputs of BiLSTM in sequence, calculate the softmax of the outputs (attention weights), multiply the attention weights with outputs of BiLSTM, then sum the outputs of the multiplication.

3.2.3. BERT Model

Finally, we investigated the recent Bidirectional Encoder Representations from Transformers (BERT) model [36], a transformer-based machine learning technique for NLP pre-training. The BERT process corrupts some tokens in the input by replacing them with MASK and then trains a model to recreate the original tokens (Figure 3). BERT uses unlabeled text to pre-train deep bidirectional representations, unlike other older language representation models, by jointly conditioning the right and left context in all layers. As a result, fine-tuning the pre-trained BERT model can be done using one extra output layer to provide state-of-the-art models for a big range of NLP tasks. In this work, we fine-tuned the BERT model to produce an Arabic dialect identification model.

In our experiments, we utilized three pre-trained BERT models as the base models to be tuned. Pre-trained BERT models cannot be used directly for dialect identification. So given an Arabic BERT model, we used the AOC training data subset to fine-tune the model where the input is a sequence of comments associated with a specific dialect and the output is compared to the target dialect for minimizing the difference to fine-tune the

weights to convert the model into a dialect identifier. Once the model is fine-tuned, it can be queried using an Arabic sentence and the model should propose a specific dialect for the input. A summary of pre-trained BERT models is the following.

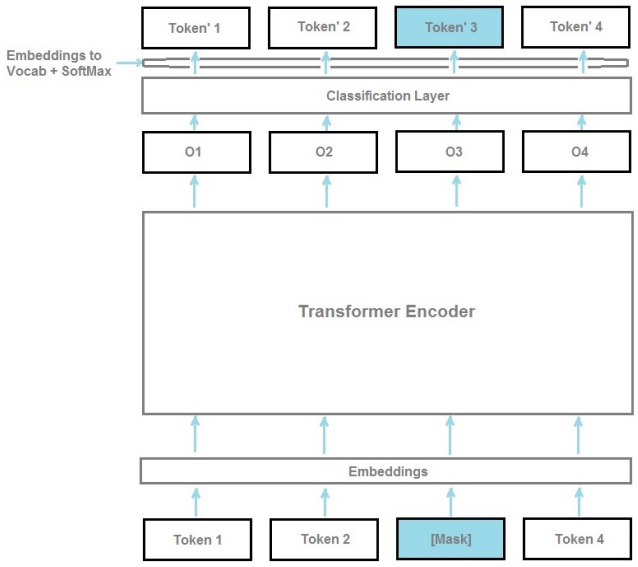

Figure 3. Bidirectional Encoder Representations from Transformers (BERT) model.

ArabicBERT

ArabicBERT was the first pre-trained BERT model for Arabic [37] that used a corpus consists of an unshuffled version of OSCAR data [32] and a relatively recent data dump from Wikipedia, which sums up to 8.2 B words and a vocabulary set of 32,000 word pieces. The corpus and the vocabulary set were not restricted to Modern Standard Arabic (MSA) and contained some dialectical Arabic too, which boosted the model's performance in terms of data from social media platforms. The main specifications of the ArabicBERT model are:

- Hidden size: 768;
- Num attention heads: 12;
- Num hidden layers: 12;
- Vocab size: 32,000.

AraBERT

AraBERT was pre-trained specifically for Arabic, hoping to replicate BERT's success with state-of-the-art results on the majority of tested Arabic natural language processing tasks [32]. It has been trained using crawling Arabic news websites, among other sources. The performance of AraBERT is better than multilingual BERT from Google (mBERT) in several downstream tasks. Specifications of the AraBERT model are summarized as follows:

- Hidden size: 768;
- Num attention heads: 12;
- Num hidden layers: 12;
- Vocab size: 64,000.

MARBERT

The third model was MARBERT [38]. MARBERT is an Arabic large-scale masked language model that was proposed this year. It targets both MSA and dialectical Arabic. It has been trained using a large dataset of Arabic tweets. MARBERT is best suited for downstream tasks involving dialectal Arabic. It also has state-of-the-art results in many of the downstream tasks in Arabic. The following are the main characteristics of the model:

- Hidden size: 768;

- Num attention heads: 12;
- Num hidden layers: 12;
- Vocab size: 100,000.

4. Experiments and Discussion

Our experiments were conducted on Intel(R) machine (Xeon(R) Gold 6230 CPU @ 2.10GHz 20*2 Cores with 1 Tesla V100 32GB GPU). We trained our models with the largest available Arabic Dialectal data set, which is AOC. AOC consists of 3 million MSA and dialectal comments, with 108k of them labeled by utilizing crowdsourcing. We used the following data splits: 80% for training (Train), 10% for validation (Dev), and 10% for the test (Test) after shuffling the 108 K comments at random. Table 1 and Figure 4 illustrate the data distribution through the different segmentations and the shared vocabulary between the dialects [13].

Table 1. Distribution of dialects in the used Arabic Online Commentary (AOC) dataset [13].

Variety	MSA	EGY	GLF	LEV	ALL
Train	50,845	10,022	16,593	9081	86,541
Dev	6357	1253	2075	1136	10,821
Test	6353	1252	2073	1133	10,812

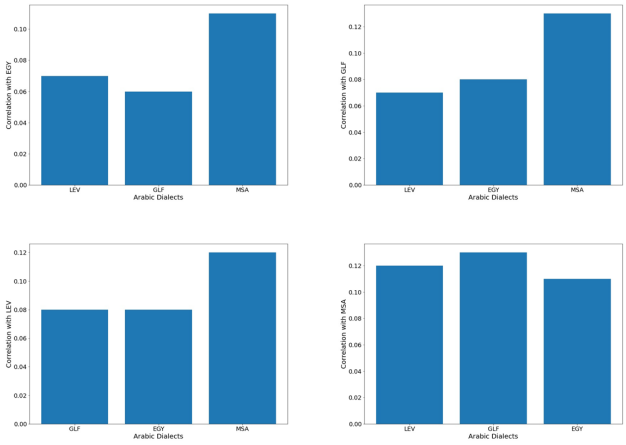

Figure 4. Distribution of shared vocabulary between the dialects within the used AOC dataset.

We performed two different classification tasks: (A) 3-way dialects, in which we estimate the difference between LEV, EGY, and GLF; and (B) 4-way variants (i.e., LEV vs. EGY vs. GLF vs. MSA).

4.1. Embeddings Types

The following are a list of the embeddings' types used in the experiments:

- AraBERT embeddings: This type uses the encodings of subwords used by the AraBERT model.
- ArabicBERT embeddings: Similarly, this type utilizes the subwords encodings of the ArabicBERT model.
- MARBERT embeddings: These are the encodings used by the MARBERT model.
- Twitter-City embeddings: 300-dimensional word vectors proposed in [39].
- AOC-based embeddings: These embeddings are extracted from a continuous bag of words model described in [13] to generate 300-dimensional word vectors.
- Random embeddings: These embeddings are generated by randomly initializing the input layer.

4.2. AOC-Based Results

Tables 2 and 3 show our results when using AOC data for training and fine-tuning for the 4-way and 3-way dialect identification using all the experimented classifiers and the different word embedding techniques. It can be observed from Table 2 that the LR classical classifier is still a good candidate for the application of dialect identification due to its simplicity and efficiency for model training. Furthermore, most of the dialect effect can be captured just from the top frequent 5000 words. Using larger word lists would not provide better performance. Due to the stochastic behavior of several processes of neural network training, we applied each neural network-based experiment three times. The results illustrated in Tables 2 and 3 that are related to neural-based experiments are the average values for each experiment. The neural networks-based models significantly improve performance with the best result achieved with the pre-trained BERT-based models using the MARBERT base and embedding, which achieved an accuracy of 85.86%.

By excluding the MSA part, which represented a dominant segment of the data, and repeating the experiments for 3-way dialect identification, the fine-tuned MARBERT-based model also achieved the best performance with an accuracy of 90.45%.

To check the stability of the proposed models, we conducted several experiments. Using our best model (MARBERT based), we checked the sensitivity of the model via changing two of the main parameters: Dropout rate and batch size. First, we fixed the batch size as 16 and applied the dropout rate values 0.0, 0.3, 0.5, and 0.8. The best accuracy was achieved using these settings with a dropout rate of 0.5. So, the value of dropout was maintained at 0.5 and batch size values of 24, 32, 48, 64, and 84 were applied. Table 4 illustrates the results of parameters sensitivity experiments.

As shown in Table 4, the accuracy values change from as low as 83.91% to as high as 86.17% when changing the values of dropout rate and batch size. Although this change in the accuracy value is not small, the whole range of values is higher than the state-of-the-art result of [13] (82.45%). However a careful selection of hyperparameters should be made to maximize the achieved gain.

Table 2. The 4 classes dialect identification results.

Embeddings Level	Embeddings Type	Model	Precision	Recall	F1 Score	Accuracy
Word (unigrams, bigrams, trigrams)	TFIDF	Logistic Regression	0.7647	0.6723	0.7019	0.7703
Word (unigrams)	TFIDF	Logistic Regression	0.7949	0.667	0.712	0.7829
Word (1k unigrams)	TFIDF	Logistic Regression	0.7785	0.6635	0.7036	0.7756
Word (5k unigrams)	TFIDF	Logistic Regression	0.7972	0.6766	0.7199	0.7879
Word (10k unigrams)	TFIDF	Logistic Regression	0.7992	0.6727	0.7175	0.7861
Word (20k unigrams)	TFIDF	Logistic Regression	0.7986	0.6693	0.7149	0.7847
Word (unigrams)	Random Embeddings	BiLSTM	0.7783	0.7179	0.7389	0.7952
Word (unigrams)	AOC Embeddings	BiLSTM	0.778	0.7038	0.7309	0.7947
Word (unigrams)	Twitter Embeddings	BiLSTM	0.8046	0.7632	0.7811	0.8344
Word (unigrams)	Random Embeddings	Attention BiLSTM	0.7814	0.7097	0.7344	0.7932
Word (unigrams)	AOC Embeddings	Attention BiLSTM	0.78	0.6842	0.7189	0.7871
Word (unigrams)	Twitter Embeddings	Attention BiLSTM	0.8001	0.7577	0.7748	0.8278
Subword	AraBERT Embeddings	AraBERT	0.8086	0.7564	0.7777	0.8364
Subword	ArabicBERT Embeddings	ArabicBERT	0.8132	0.7677	0.787	0.8396
Subword	MARBERT Embeddings	MARBERT	0.8402	0.8000	0.8157	0.8586

Figure 5 illustrates confusion matrices of MARBERT-based model result errors. The right confusion matrix shows the four dialects classification errors while the left one presents the errors of the three dialect classification task. As we can see, the confusion of dialects between each other in the 3-dialect experiment is relatively low. However in the 4-dialect experiment, there is a notable perplexity between MSA & LEV (19%) and MSA & GLF (18%). This is aligned with the fact that there is a considerable overlap of vocabulary between MSA & LEV and MSA & GLF as we have seen in Figure 4.

Table 3. The 3 classes dialect identification results.

Embeddings Level	Embeddings Type	Model	Precision	Recall	F1 Score	Accuracy
Word (unigrams)	TFIDF	Logistic Regression	0.87	0.83	0.85	0.853
Word (1k unigrams)	TFIDF	Logistic Regression	0.86	0.82	0.84	0.8423
Word (5k unigrams)	TFIDF	Logistic Regression	0.87	0.83	0.85	0.8553
Word (10k unigrams)	TFIDF	Logistic Regression	0.88	0.84	0.85	0.8568
Word (20k unigrams)	TFIDF	Logistic Regression	0.87	0.83	0.85	0.8548
Word (unigrams, bigrams, trigrams)	TFIDF	Logistic Regression	0.83	0.82	0.83	0.8315
Word (unigrams)	Random Embeddings	BiLSTM	0.8703	0.8531	0.8605	0.8659
Word (unigrams)	AOC Embeddings	BiLSTM	0.8522	0.8393	0.8449	0.8513
Word (unigrams)	Twitter Embeddings	BiLSTM	0.8742	0.8669	0.8703	0.876
Word (unigrams)	Random Embeddings	Attention BiLSTM	0.8568	0.8507	0.8533	0.8577
Word (unigrams)	AOC Embeddings	Attention BiLSTM	0.847	0.8375	0.8418	0.8473
Word (unigrams)	Twitter Embeddings	Attention BiLSTM	0.8792	0.8697	0.874	0.8793
Word (unigrams)	Random Embeddings	Self Attention	0.8325	0.8147	0.8221	0.8294
Word (unigrams)	Twitter Embeddings	Self Attention	0.8467	0.8221	0.832	0.8423
Word (unigrams)	AOC Embeddings	Self Attention	0.824	0.7935	0.8048	0.8158
Word (unigrams)	Random Embeddings	Transformer Encoder	0.8619	0.8536	0.8574	0.8585
Word (unigrams)	Twitter Embeddings	Transformer Encoder	0.8563	0.8462	0.8507	0.8573
Word (unigrams)	AOC Embeddings	Transformer Encoder	0.803	0.7753	0.7857	0.7995
Subword	AraBERT Embeddings	AraBERT	0.8802	0.8613	0.8692	0.8749
Subword	ArabicBERT Embeddings	ArabicBERT	0.8852	0.8654	0.8739	0.8739
Subword	MARBERT Embeddings	MARBERT	0.9039	0.8993	0.9011	0.9045

Table 4. The 4 classes dialect identification results of our best proposed model for a wide range of dropout rate and batch size values.

Dropout Rate	Batch Size	Precision	Recall	F1 Score	Accuracy
0.0	16	0.8198	0.7611	0.7872	0.8424
0.3	16	0.8166	0.7679	0.7882	0.8391
0.5	16	0.8204	0.7638	0.7889	0.8429
0.8	16	0.8173	0.7584	0.7844	0.8396
0.5	24	0.8454	0.8026	0.8177	0.8578
0.5	32	0.8504	0.7946	0.8156	0.8588
0.5	48	0.8423	0.7938	0.8053	0.8503
0.5	64	0.8338	0.8020	0.8066	0.8500
0.5	84	0.8511	0.8014	0.8203	0.8617

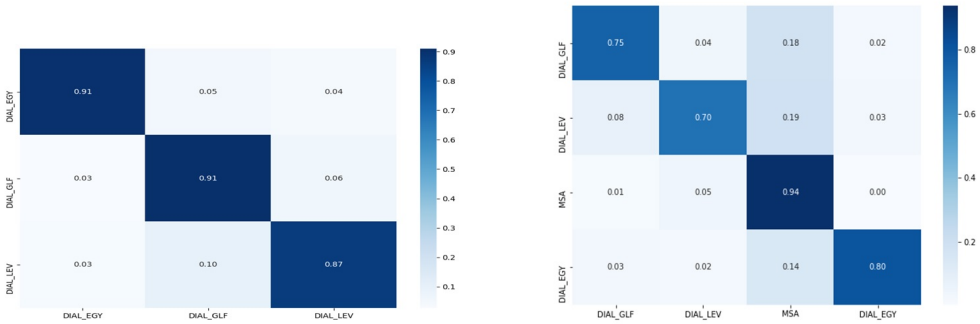

Figure 5. Analysis of MarBERT-based model results. (**Left**): The confusion matrix for 3-dialects classification. (**Right**): The confusion matrix for 4-dialects classification.

4.3. COVID-19 Twitter Data Results

The focus in the dataset we utilized which is described in [7] was on hashtags used in Saudi Arabia, although they might be used in Arabic-speaking countries outside of Saudi Arabia. Saudi Arabia is one of the countries with the highest amount of Twitter users among the country's online population [40,41]. Saudi Arabia also generates 40 percent of

all tweets all over Arab countries [42]. The used dataset has several sections including conversations discussing the precautionary measures governments have applied, conversations discussing social solidarity, and conversations supporting the governments' decisions. Moreover, the dataset has data from three Saudi official accounts. The amount of tweets in the dataset is around 3.8 million. We applied the BERT-based dialect recognition model on this dataset and selected the tweets that achieved a 90% confidence score (based on the SoftMax score), which provided 1.76 million tweets. We randomly selected 2500 tweets from each region set and made a manual review for them. Table 5 shows the distribution of the labeled region dialects for the 1.76 million tweets and the accuracy results for the randomly selected set.

Table 5. COVID-19 labeled tweets and dialect recognition result.

Region	Number of Tweets	Accuracy
MSA	1,373,869	97.7%
Gulf	222,083	97.6%
Egyptian	139,143	92.8%
Levantine	34,037	99.6%
Total	1,769,132	97.36%

As shown in Table 5, our model achieved significant results for region detection with a weighted average accuracy of 97.36%.

4.4. Discussion

While our models used AOC data for training and testing, we can compare our results to state-of-the-art models that used the same data. As mentioned in the Related Work section, the authors of [13] proposed the AOC data-based state-of-the-art models. They reported accuracy of 87.41% and 82.45% for 3-way and 4-way classification tasks respectively. According to these results, we have over 3% absolute gain in the 3-way classification task, and over 3.4% absolute gain when dealing with the 4-way classification task. When applied to the Twitter COVID-19 dataset, the proposed model performance was notable to the level that makes it an important knowledge source for applying geo-based analysis of pandemic disasters text data.

Regarding the limitations of the proposed work, it is worth mentioning that the Twitter dataset we used for verification is domain-specific. So, it is expected to get a higher accuracy above the result achieved in the AOC-based test. In addition, the current versions of the suggested models do not support some Arabic dialects like Maghrebi. However, the dialects we have addressed in the experiments cover about half of the Arab population along with the modern standard Arabic that is common among all Arabs.

5. Conclusions

In this work, we introduced an enhanced method for Arabic language dialect identification. We investigated several techniques such as logistic regression and recent advanced deep learning-based classifiers. We achieved the best performance using the BERT-based model that provided 90.45% accuracy on the Arabic Online Commentary (AOC) dataset. When we applied the BERT-based dialect recognition model for identifying the source region of COVID-19 related tweets, we managed to recognize the source region for 1.76 million Arabic tweets. After randomly sampling 2500 tweets from each region set and checking them manually, we got a weighted average accuracy of 97.36%. The tweets from January to April 2020 were used for experimentation, which was the starting critical period of the virus spread. The high accuracies of the results show that the proposed model can be used as a powerful tool for getting geo-based insights for disaster-related social data. We plan to expand dialect recognition to the country level in future work, which is challenging due to the overlap of the local dialects in neighboring countries.

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

Funding: The authors are grateful to the Saudi Arabian Ministry of Education's Deputyship for Research and Innovation for supporting this research through project number 523.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: [7,12].

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

References

1. World Health Organization. *Corona Virus Disease 2019 (COVID-19): Situation Report*; World Health Organization: Geneva, Switzerland, 2020; Volume 83.
2. World Health Organization. *COVID-19 Weekly Epidemiological Update*; World Health Organization: Geneva, Switzerland, 2021.
3. Sakaki, T.; Okazaki, M.; Matsuo, Y. Earthquake shakes twitter users: Real-time event detection by social sensors. In Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC, USA, 26–30 April 2010; pp. 851–860.
4. Palen, L.; Hughes, A.L. Social media in disaster communication. In *Handbook of Disaster Research*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 497–518.
5. Karami, A.; Shah, V.; Vaezi, R.; Bansal, A. Twitter speaks: A case of national disaster situational awareness. *J. Inf. Sci.* **2020**, *46*, 313–324. [CrossRef]
6. Hariharan, K.; Lobo, A.; Deshmukh, S. Hybrid Approach for Effective Disaster Management Using Twitter Data and Image-Based Analysis. In Proceedings of the 2021 International Conference on Communication information and Computing Technology (ICCICT), Mumbai, India, 25–27 June 2021.
7. Addawood, A. Coronavirus: Public Arabic Twitter Data Set. 2020. Available online: https://openreview.net/forum?id=ZxjFAfD0pSy (accessed on 22 November 2021).
8. Imène, G.; Azouaou, F. Arabic dialect identification with an unsupervised learning (based on a lexicon) application case: Algerian dialect. In Proceedings of the 2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES), Paris, France, 24–26 August 2016.
9. Habash, N.Y. Introduction to Arabic natural language processing. *Synth. Lect. Hum. Lang. Technol.* **2010**, *3*, 1–187. [CrossRef]
10. Abdul-Mageed, M. Subjectivity and Sentiment Analysis of Arabic as a Morphologically-Rich Language. Ph.D. Thesis, Indiana University, Bloomington, IN, USA, 2015.
11. Abdul-Mageed, M.; Buffone, A.; Peng, H.; Eichstaedt, J.C.; Ungar, L.H. Recognizing pathogenic empathy in social media. In *ICWSM*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 448–451.
12. Zaidan, O.F.; Callison-Burch, C. The arabic online commentary dataset: An annotated dataset of informal Arabic with high dialectal content. In Proceedings of the ACL, Portland, OR, USA, 19–24 June 2011; pp. 37–41.
13. Elaraby, M.; Abdul-Mageed, M. Deep models for Arabic dialect identification on benchmarked data. In Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018), Santa Fe, NM, USA, 20 August 2018.
14. Belinkov, Y.; Glass, J. A character-level convolutional neural network for distinguishing similar languages and dialects. *arXiv* **2016**, arXiv:1609.07568.
15. Shon, S.; Ali, A.; Glass, J. Mit-qcri arabic dialect identification system for the 2017 multi-genre broadcast challenge. *arXiv* **2017**, arXiv:1709.00387.
16. Shon, S.; Ali, A.; Glass, J. Convolutional neural networks and language embeddings for end-to-end dialect recognition. *arXiv* **2018**, arXiv:1803.04567.
17. Elfardy, H.; Diab, M. Sentence level dialect identification in Arabic. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Sofia, Bulgaria, 4–9 August 2013; Volume 2, pp. 456–461.
18. Elfardy, H.; Al-Badrashiny, M.; Diab, M. Aida: Identifying code switching in informal arabic text. In Proceedings of the First Workshop on Computational Approaches to Code Switching, Doha, Qatar, 25 October 2014; pp. 94–101.
19. Zaidan, O.F.; Callison-Burch, C. Arabic dialect identification. *Comput. Linguist.* **2014**, *40*, 171–202. [CrossRef]
20. Cotterell, R.; Callison-Burch, C. A multi-dialect, multi-genre corpus of informal written arabic. In Proceedings of the LREC, Reykjavik, Iceland, 26–31 May 2014; pp. 241–245.
21. Darwish, K.; Sajjad, H.; Mubarak, H. Verifiably effective arabic dialect identification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1465–1468.

22. Mousa, A. Deep Identification of Arabic Dialects. Informatics Institute. Bachelor's Thesis, Karlsruhe Institute of Technology, Karlsruhe, Germany, 2021.
23. Elfardy, H.; Al-Badrashiny, M.; Diab, M. Code switch point detection in Arabic. In *International Conference on Application of Natural Language to Information Systems*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 412–416.
24. Pasha, A.; Al-Badrashiny, M.; Diab, M.T.; El Kholy, A.; Eskander, R.; Habash, N. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In Proceedings of the Lrec, Reykjavik, Iceland, 26–31 May 2014; pp. 1094–1101.
25. Ragab, A.; Seelawi, H.; Samir, M.; Mattar, A.; Al-Bataineh, H.; Zaghloul, M.; Mustafa, A.; Talafha, B.; Freihat, A.A.; Al-Natsheh, H. Mawdoo3 AI at MADAR Shared Task: Arabic Fine-Grained Dialect Identification with Ensemble Learning. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, Florence, Italy, 28 July–2 August 2019; pp. 244–248.
26. Althobaiti, J.M. Automatic Arabic dialect identification systems for written texts: A survey. *arXiv* **2020**, arXiv:2009.12622.
27. Ghoul, D.; Lejeune, G. MICHAEL: Mining Character-level Patterns for Arabic Dialect Identification (MADAR Challenge). In Proceedings of the Fourth Arabic Natural Language Processing Workshop, Florence, Italy, 28 July–2 August 2019; pp. 229–233.
28. Přibáň, P.; Taylor, S. ZCU-NLP at MADAR 2019: Recognizing Arabic Dialects. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, Florence, Italy, 28 July–2 August 2019; pp. 208–213.
29. Harrat, S.; Meftouh, K.; Abidi, K.; Smaïli, K. Automatic identification methods on a corpus of twenty five fine-grained arabic dialects. In *International Conference on Arabic Language Processing*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 79–92.
30. Huang, F. Improved arabic dialect classification with social media data. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 2118–2126.
31. Aizawa, A. An information-theoretic perspective of tf-idf measures. *Inf. Process. Manag.* **2003**, *39*, 45–65. [CrossRef]
32. Antoun, W.; Baly, F.; Hajj, H. AraBERT: Transformer-based model for Arabic language understanding. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, Marseille, France, 11–16 May 2020; pp. 9–15.
33. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
34. Chorowski, J.K.; Bahdanau, D.; Serdyuk, D.; Cho, K.; Bengio, Y. Attention based models for speech recognition. *arXiv* **2015**, arXiv:1506.07503.
35. Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; Xu, B. Attention-based bidirectional long short-term memory networks for relation classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Berlin, Germany, 7–12 August 2016; pp. 207–212.
36. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
37. 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.
38. Abdul-Mageed, M.; Elmadany, A.; Nagoudi, E. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. *arXiv* **2020**, arXiv:2101.01785.
39. Abdul-Mageed, M.; Alhuzali, H.; Elaraby, M. You tweet what you speak: A city level dataset of arabic dialects. In Proceedings of the LREC, Miyazaki, Japan, 7–12 May 2018; pp. 3653–3659.
40. Clement, J. Countries with Most Twitter Users 2020. 2020. Available online: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/ (accessed on 22 November 2021).
41. Puri-Mirza, A. Saudi Arabia: Number of Internet Users 2023. 2019. Available online: https://www.statista.com/statistics/462959/internet-users-saudi-arabia/ (accessed on 22 November 2021).
42. Mourtada, R.; Salem, F. Citizen engagement and public services in the arab world: The potential of social media. In *Arab Social Media Report Series*, 6th ed.; Mohammed Bin Rashid School of Government, SSRN: Dubai, United Arab Emirates, 2014.