IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 26, No. 5, October, 2021, pp. 469-475
Journal homepage: http://iieta.org/journals/isi

Impact of Using Bidirectional Encoder Representations from Transformers (BERT) Models for Arabic Dialogue Acts Identification

Alaa Joukhadar[1*], Nada Ghneim[2], Ghaida Rebdawi[1]

[1] Information Technology Department, Higher Institute for Applied Sciences and Technology, Damascus 31983, Syria
[2] Faculty of Informatics & Communication Engineering, Arab International University, Damascus 16180, Syria

Corresponding Author Email: alaa.joukhadar@hiast.edu.sy

https://doi.org/10.18280/isi.260506

Received: 10 September 2021
Accepted: 25 October 2021

Keywords:
AraBERT models, Arabic language, dialogue acts identifications, user intents identification

ABSTRACT

In Human-Computer dialogue systems, the correct identification of the intent underlying a speaker's utterance is crucial to the success of a dialogue. Several researches have studied the Dialogue Act Classification (DAC) task to identify *Dialogue Acts (DA)* for different languages. Recently, the emergence of Bidirectional Encoder Representations from Transformers (BERT) models, enabled establishing state-of-the-art results for a variety of natural language processing tasks in different languages. Very few researches have been done in the Arabic Dialogue acts identification task. The BERT representation model has not been studied yet in Arabic Dialogue acts detection task. In this paper, we propose a model using BERT language representation to identify Arabic Dialogue Acts. We explore the impact of using different BERT models: AraBERT Original (v0.1, v1), AraBERT Base (v0.2, and v2) and AraBERT Large (v0.2, and v2), which are pretrained on different Arabic corpora (different in size, morphological segmentation, language model window, …). The comparison was performed on two available Arabic datasets. Using AraBERTv0.2-base model for dialogue representations outperformed all other pretrained models. Moreover, we compared the performance of AraBERTv0.2-base model to the state-of-the-art approaches applied on the two datasets. The comparison showed that this representation model outperformed the performance both state-of-the-art models.

1. INTRODUCTION

Human-Computer dialogue systems provide a natural language based interface between humans and computers. They have a broad range of applications such as: Intelligent Tutoring systems [1], e-Health Applications [2], Argumentation Detection [3], Natural Language Generation [4], etc. The Human-Computer dialogue system can be composed of different components: The speech signal of the user is converted to a sequence of words by the Automatic Speech Recognition (ASR) system. This words sequence is then handled by the Spoken Language Understanding (SLU) module to find a meaningful interpretation of the user's intentions. The Dialogue Management (DM) module chooses the communicative action the system must perform. Finally, the Natural Language Generation (NLG) module creates a sentence that best expresses the system's intention. The correct interpretation of the *intent* (also named as *speech or dialogue acts*) of each speaker's utterance plays a crucial role in the success of the whole dialogue, as it is essential to help the Dialogue Manager component determining the next action of the system, such as answering with the correct information when the user is asking a question, acknowledging the reception of an answer, or keeping quiet when the user is just giving a simple comment.

The available dialogue acts annotated resources are still modest compared to other language resources, with annotations that are only partly compatible with each other. Dialogue acts annotations usually differ with respect to the domain (restaurants, hotels, …), and by their granularity level. Thus, such annotated corpora are not easy to reuse for other purposes neither to apply to domains other than they were originally developed for. Different approaches have been used to recognize speech acts in dialogues, starting from rule-based systems, to machine learning and deep learning methods. New language representation models have recently been developed to overcome the lack of sufficient training data. Bidirectional Encoder Representations from Transformers (BERT), a newly introduced language representation model that provides pretrained deep bidirectional representations of vast unlabeled data, had drastically improved the efficiency of NLP applications. Very few researches have been done in the Arabic Dialogue acts identification task. In this paper, we focus on the Dialogue acts recognition in Arabic language. But, since Arabs do not use the Standard Arabic language to interact, but rather different dialects generally classified by regions (North African, Levantine, Egyptian, …) [5], we will study the dialectal Arabic dialogue acts. The main contributions of this work are as follows:

(i) To the best of our knowledge, it is the first study that apply different BERT language representation models in the Arabic Dialogue Act identification task, and compare their effectiveness. These models are: AraBERT Original (v0.1, v1), AraBERT Base (v0.2, and v2) and AraBERT Large (v0.2, and v2).

(ii) Extensive experiments and analysis on two datasets (in Levantine and Egyptian dialects) verify the effectiveness of our proposed model and show that it outperforms other state-of-the-art models.

The remainder of this article is organized as follows:

Section 2 presents a brief description of the related works. The proposed approach has been discussed in Section 3. The experimental results and the corresponding analysis are given in Section 4. Finally, the concluding remarks and the directions for future work are discussed in Section 5.

2. RELATED WORKS

In Human-Machine Interaction dialogue systems, the Dialogue Acts recognition is considered an important component. The research in this area has made great progress during the last few years. Kumar et al. [6] was based on a hierarchical recurrent neural network that learns representations at word, utterance, and conversation levels. The conversation level representations consider all previous utterances and their dialogue acts. They validated their approach on two datasets: Switchboard (SwDA) and Meeting Recorder Dialogue Act (MRDA). The performance outperformed the state-of-the-art methods by 2.2% and 4.1% absolute points, respectively. a simple RNN has been used to model the previous utterances context. They evaluated their model on the Switchboard dataset and achieved an accuracy of 77.34% [7]. Lee and Dernoncourt [8] have also presented a model based on RNN and CNN networks that incorporates the preceding short texts. Their model was validated on three different datasets (DSTC 4, MRDA, and SwDA), and achieved state-of-the-art results. Chen et al. [9] proposed a 2-steps approach based on the CRF-Attentive Structured Network. The approach outperformed several state-of-the-art solutions on SwDA and MRDA datasets. An improved dynamic memory network with hierarchical pyramidal utterance encoder has been proposed by Wan et al. [10]. Experiments showed that the model was robust and achieved better performance compared with some state-of-the-art baselines. Dai et al. [11] proposed a hierarchical model to capture intra-sentence and inter-sentence information based on self-attention. Their model was tested on two datasets: SwDA and DailyDialog, and achieved promising performance with an accuracy of 80.34% and 85.81% respectively.

With the introduction of Bidirectional Encoder Representations from Transformers (BERT), lots of NLP tasks started outperforming other models. Saha et al. [12] presented a novel model for the identification of speech acts in Twitter on top of BERT. They introduced a BERT-extended classifier for the task, with a model based on calculating the attention weights over the token representations of a sequence obtained from the pre-trained BERT model. The proposed model was evaluated on an open-access data set released by Saha et al. [13] and outperformed the state-of-the-art approach, with an accuracy of 75.97%. Then Saha et al. introduced [14] the BERT-Caps model, which is built on top of BERT. The new model takes into consideration both feature optimizations from BERT and capsule layer to better learn the traits and attributes. Their results outperformed several strong baselines and state-of-the-art approaches with an accuracy of 77.52%. Saha et al. [15] studied the role of sentiment and emotion in speech act classification in Twitter. They proposed a Dyadic Attention Mechanism based multi-modal (emojis and text), adversarial multi-task framework for joint optimization of Tweet Acts, sentiment and emotions. Their experiments on EmoTA tweets acts dataset, showed that the proposed framework boosts the classification performance by benefitting from the Sentiment and Emotion Analysis.

Concerning non-English languages, different researches focused on multilingual domain. Cerisara et al. [16] explored in recurrent models to capture the sequence of words within sentences, and studied the impact of using pre-trained word embedding representations on the recognition task. The model was tested on English, French and Czech, with a consistent performance across the three languages, and a comparable performance to the state-of-the-art results in English. For Persian language, a dictionary-based statistical technique was proposed for speech acts recognition [17]. Authors used lexical, syntactic, semantic, and surface features to detect seven classes of speech acts. They evaluated their proposed technique using four classification methods :Random Forest RF, Support Vector Machine SVM, Naive Bayes NB, and K-Nearest Neighbors KNN. The best classification accuracy was obtained using RF and SVM.

Recently, Arabic speech acts classification task started to show preliminary initiatives. Early works were dedicated to Modern Standard Arabic (MSA). Shala et al. [18] applied different machine learning classifiers (SVM, NB and Decision Trees) to classify Arabic discourse speech acts using a dataset of about 400 MSA utterances collected from newspapers. Sherkawi et al. [19] presented a bootstrapped rule-based model to detect MSA Arabic Speech Act types. The studied speech acts types were: Affirmation, Hope, Condition, Praise, Dispraise, Negation, Confirmation, Interrogation, Imperative, Forbidding, Wishing, Vocative, Prompting, Rebuke, Exclamation, Swear. The model was tested on a corpus of about 1500 MSA sentences. Sherkawi at al. [20] proposed a machine learning approach based on surface features, cue words and contextual information to recognize MSA Arabic speech acts. Authors compared the results of Decision Trees, Naïve Bayes, Neural Networks and SVM algorithms on their corpus of 1500 MSA sentences. The Decision Tree algorithm had the best results. Other studies focused on the dialectal language. For Tunisian Dialect Graja et al. [21] used the TuDiCoI corpus (12182 utterances) to develop a discriminative algorithm based on CRF to semantically label spoken turns which are not segmented into utterances. For Egyptian Dialect, Elmadany et al. [22] proposed a machine learning approach based on multi-classes hierarchical structure to classify dialogue acts using the JANA corpus (4725 utterances). The model attained an average F-measure scores of 91.2%. Algotiml et al. [23] proposed a speech act classification model for Twitter asynchronous conversations, where they studied several machine learning methods including SVM and deep neural networks. They applied the proposed methods on the ArSAS tweets dataset [24]. The results showed that deep learning methods had better performance compared to SVMs, where Bi-LSTM achieved an accuracy of 87.5% and a macro-averaged F1 score 61.5%. Joukhadar et al. [25] compared various machine learning algorithms with different features being used to detect the correct speech act categories. They compared the results of the proposed models on a hand-crafted corpus of 873 sentences in the restaurant's orders and airline ticketing domain in Levantine Arabic Dialect. The best result was given by SVM algorithm with 2-gram word features.

3. OUR APPROACH

The first subsection describes the datasets that were used in this work. The following two subsections describe the models

implemented for comparison, with their setups and hyperparameters used.

3.1 Dataset

Two different datasets were used to evaluate the models: ArSAS released by Elmadany et al. [24], and the one used by Joukhadar et al. [25], which we will call LevInt.

3.1.1 ArSAS dataset

ArSAS is an open-access data set, which contains 21,081 Arabic tweets in different Arabic dialects. It is annotated by six speech act classes, defined by Elmadany et al. [24] as:

i. Assertion: The user declares some proposition such as stating, claiming, reporting, or announcing.
ii. Recommendation: the user recommends something.
iii. Expression: The user expresses some psychological state such as thanking, apologizing, or congratulating.
iv. Question: The user asks a question such as why, what, or confirmation.
v. Request: The user asks for something such as ordering, requesting, demanding, or begging.
vi. Miscellaneous: The user is committed to some future action such as promising or offering.

The tweets in the corpus cover 20 topics including long-standing topics, events and entities (celebrities or organizations). Table 1 shows the number of tweets for each of these types present in the data set.

Table 1. The number of tweets in each Dialogue act

Ass.	Rec.	Exp.	Que.	Req.	Misc.
8,221	107	11,745	751	180	60

Table 2 shows an example tweet for each of the tweet acts as given by Elmadany et al. [24].

Table 2. An example tweet for each of the tweet acts

Speech Act	Examples
Ass.	#الشروق: السيسي: كلي فخر واعتزاز بالنخبة المتميزة المشاركة في شباب العالم #Sunrise: El-Sisi: I am proud of all the elite who are contributing in the world cup forum
Exp.	أشعر أن الربيع العربي إشعاع من الحرية I feel that Arab revolutions are a radiation of freedom
Rec.	الكرة الايطالية تحتاج لتركي آل الشيخ Italian football needs Turkey Al-Saikh
Que.	ما هو رايكم في مقاطعة السعودية لدولة قطر What is your opinion in Saudi banning Qatar?
Req.	بعد منتدى شباب العالم أطالب بعمل منتدى للفساد لإظهار الحقائق After the world cup forum, I request to do a forum to reveal truths
Mis.	وليد سليمان ممكن يلعب مكان محمد صلاح وممكن يلعب مكان عبد الشافي Walid Sulaiman can play instead of Muhammad Salah or possibly instead of Abdul-Shafy

3.1.2 LevInt dataset

This dataset contains a set of 873 Levantine Arabic sentences that were manually tagged from Restaurants Orders and Airline Ticketing domains. The dataset adopted a taxonomy of 8 speech acts: (Greeting, Goodbye, Thanks, Confirm, Negate, Ask_repeat, Ask_for_alt, and Apology) that are mostly used in restaurant orders and airline ticketing. Although the dataset is small, but to our knowledge, it's the only balanced dataset in Levantine dialect. Table 3 shows the different Dialogue acts taxonomy as given by Joukhadar et al. [25].

Table 3. The distribution dialogue acts in LevInt dataset with an example utterance for each

Dialogue Act	%	Example Utterance
Greeting	12.9	مرحبا كيفك شو أخبارك؟ Hello how are you what are you up to?
Goodbye	11.0	وعليكم السلام الله معك Peace be upon you, goodbye
Thanks	13.0	شكرا كتير Thanks a lot.
Confirm	13.6	أي أكيد Yes of course
Negate	11.8	لا ما بدي No, I don't want it
Ask_repeat	12.7	ممكن تعيد شو قلت Can you repeat what you said
Ask_for_alt	12.6	شو في عندك غير خيارات What other options do you have?
Apology	12.0	آسف Sorry

4. MODELS

Our proposed system is based on AraBERT, a pre-trained BERT transformer model (a stacked Bidirectional Transformer Encoder) for the Arabic language [26]. The basic idea of BERT is to pre-train deep bidirectional representations from unlabeled text including context from both directions then fine-tune all the output parameters on the studied task.

Following the original BERT pre-training objective, AraBERT employs the Masked Language Modelling (MLM) and the Next Sentence Prediction (NSP) tasks.

In this work, we explore the different pre-trained versions of the Arabic-specific BERT provided by Antoun et al. [26], and fine-tune them for Arabic Dialogue Acts Recognition task. The six models are: AraBERT Original (v0.1, v1), AraBERT Base (v0.2, and v2) and AraBERT Large (v0.2, and v2).

The AraBERT Original models were trained on about 23GB of Arabic text, collected from Arabic Wikipedia and news articles from different media in different Arab regions, and therefore can be representative of a wide range of topics discussed in the Arab world. This corpus contains about 77M sentences with 2.7B words. Both AraBERT Base and Large models were trained on a larger corpus of about 77GB of Arabic text, which approximately corresponds to 200M sentences with 8.6B words.

Due to the complex concatenative system of Arabic language, it has a lexical sparsity issue, where words can have different forms and share the same meaning. For example, the definite article "الـ", is always concatenated as a prefix to the next word, and never appear independently. Thus, when using an ordinary tokenizer, tokens will appear twice (with and without "الـ") which implies a huge unnecessary redundancy. To avoid this issue, both v1 and v2 of AraBERT uses pre-segmented text where words are segmented into stems, prefixes and suffixes using the Farasa Segmenter (https://farasa.qcri.org/segmentation/). For instance, "اللغة -

Alloga" becomes ة + لغ + ال " - Al+ log +a". Farasa Segmenter is an SVM[rank]-based segmenter that uses a variety of features and lexicons to rank different possible segmentations of a word. The features used account for: likelihood of stems, prefixes, suffixes, and their combination; presence in lexicons containing valid stems and named entities; and underlying stem templates. Table 4 shows the different parameters for each model version.

Table 4. The different parameters for each BERT model version

Model	Size (MB)	Params (M)	Farasa Segments	Size (GB)	#Sentences (M)	#Words (B)	#Hidden units
v0.1-orig	543MB	136M	No	23	77	2.7	768
v1-orig	543MB	136M	Yes	23	77	2.7	768
v0.2-base	543MB	136M	No	77	200	8.6	768
v2-base	543MB	136M	Yes	77	200	8.6	768
v0.2-large	1.38G	371M	No	77	200	8.6	1024
v2-large	1.38G	371M	Yes	77	200	8.6	1024

The first step of the process consists in tokenizing and representing the input sentence using the studied AraBERT model. Diacritics and extensions were removed, but English characters have been retained, as it is common to mention named entities, scientific or technical terms in its original language. Wordpeice was used for tokenization step. If the studied AraBERT model uses morphologically segmented words, we used Farasa for pre-segmentation step. Then, a fully connected neural network is applied on the output of the AraBERT model to classify the sentences according to their Dialogue acts. Figure 1 presents the architecture of the model.

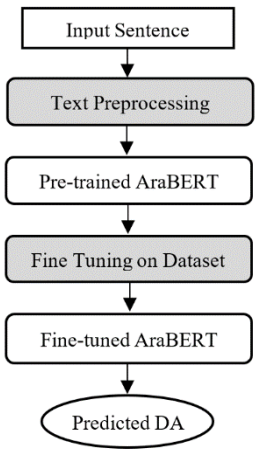

Figure 1. The architecture of the model

5. SETUP AND HYPER-PARAMETERS

In the case of ArSAS, the dataset is very unbalanced. Figure 2 shows the number of samples in each class of this dataset. We can notice that, for example, the class "Expression" has 195 times more data than the class "Miscellaneous".

In order to address the problem of the dataset imbalance, and based on Random Undersampling (RUS), and Random Oversampling (ROS), we applied the *Imbalanced Dataset Sampler* (https://github.com/ufoym/imbalanced-dataset-sampler) provided by PyTorch, that is able to rebalance the class distributions when sampling from the imbalanced dataset, while estimating the sampling weights automatically.

In the case of LevInt, the dataset is already balanced, thus no resampling process was needed.

In our experiments on the different Arabic BERT models, we relied on the implementation provided by Hugging Face's Transformers library [27]. We used the provided Auto Model for Sequence Classification, which matches each model to the proper implementation. The hyper-parameters used in our final models are presented in Table 5.

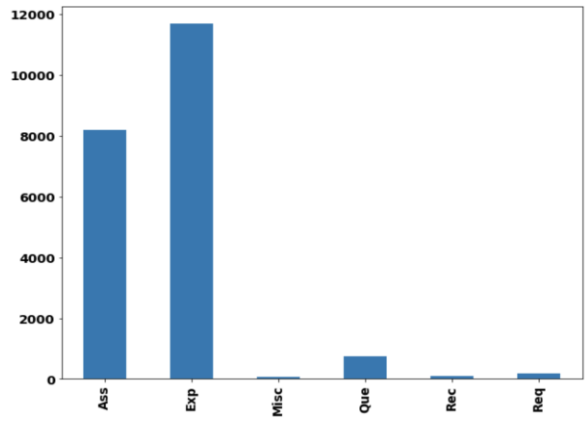

Figure 2. The number of samples in each class of ArSAS dataset

Table 5. The hyper-parameters we used for each model

Hyper-parameters	Choice
Optimizer	ADAM
Optimizer learning rate	2e-5
Learning rate	2e-5
Batch size	16
Max sentence length	360
Epochs	3

6. RESULTS

Different experiments were performed to evaluate the different Arabert model. The experiments on the ArSAS dataset and the corresponding results are presented in the first subsection, followed by the experiments on the LevInt dataset.

The performance was evaluated through various measures including: Accuracy, Macro and Weighted average of Precision, Recall and F1-measure. The classification accuracy is the total number of correct predictions divided by the total number of predictions made for a dataset. Let TP the true positive samples, FP the False positive samples, and FN the false negative samples. The precision, recall and F1 measure are given by the following formula:

$$Precision\ P = TP/(TP + FP)$$

$$Recall\ R = TP/(TP + FN)$$

$$F1\ Measure = 2 * PR/(P + R)$$

A Macro-average compute the metric independently for each class and then take the average, hence treating all classes equally, whereas a Weighted average calculates metrics for each class, and finds their average, weighted by support (the number of true instances for each class).

6.1 Model effectiveness on ArSAS dataset

As the ArSAS dataset was highly imbalanced, a resampling step was conducted before the classification. The dataset was divided into 80 % for training data and 20% for test data. Table 6 shows the results achieved by the different models on the dataset.

We remark that all models have achieved good classification results. The AraBERTv0.2-base and AraBERTv0.2-large models outperform the rest with an accuracy of 89%. The best classification performance of all the models is obtained using the AraBERTv0.2-base pre-trained model, based on Precision, Recall and F1-scores calculated using Macro and Weighted average.

To understand the results, we looked into the detailed classification report for each class. Table 7 shows the detailed classification report using the weighted average.

Table 6. Comparison between different BERT-Based models on ArSAS, in terms of Accuracy, Macro and Weighted average of Precision, Recall, and F-measure

Model	Acc.	P.	R.	F1	P.	R.	F1
		Macro Avg.			Weighted Avg.		
v0.1-Orig	0.88	0.53	0.49	0.51	0.88	0.88	0.88
v1-Orig	0.88	0.57	**0.53**	0.55	0.88	0.88	0.88
v0.2-Base	**0.89**	**0.58**	**0.53**	**0.55**	**0.89**	**0.89**	**0.89**
v2-Base	0.88	0.57	0.50	0.52	0.88	0.88	0.88
v0.2-Large	0.89	0.53	0.50	0.51	0.89	0.89	0.88
v2-Large	0.88	0.55	0.50	0.52	0.88	0.88	0.88

Table 7. The detailed classification report on ArSAS for each class (arabert-v02 model)

	Precision	Recall	F1-score	Support
Ass.	0.92	0.88	0.90	1560
Exp.	0.89	0.93	0.91	2212
Misc.	0.20	0.17	0.18	12
Que.	0.69	0.60	0.64	141
Rec.	0.29	0.20	0.24	20
Req.	0.47	0.41	0.44	34
Accuracy			0.89	3979
Macro avg	0.58	0.53	0.55	3979
Weighted avg	0.89	0.89	0.89	3979

The Expression Dialogue act has the best results with a precision, recall and F1 measure of 89%, 93%, and 91% respectively. It is followed by the Assertion with a precision, recall and F1 measure of 92%, 88%, and 90% respectively.

Although the overall performance is relatively good with an accuracy of 89% and a weighted F1-Measure of 89%, some classes have very poor performances. It could be explained by the high imbalance of the dataset classes. The Miscellaneous Dialogue act, for example, which is the smallest Dialogue act set, has a precision, recall and F1 measure of 20%, 17%, and 18% respectively, followed by the Recommendation Dialogue act, which has comparable results.

On the other hand, the Question and Request Dialogue acts have better, yet not good, results with an F1 measure of 64%, and 44% respectively (as they are still extremely small compared to Expression and Assertion sets).

Moreover, we further analysed the confusion matrix to determine the correct/incorrect labels corresponding to the sentences. Table 8 shows the confusion matrix of our best model (arabert-v02).

Table 8. The confusion matrix of our best model (Arabert-v02)

	Ass.	Exp.	Misc.	Que.	Rec.	Req.
Ass. (1560)	**1377**	160	6	10	1	6
Exp. (2212)	110	**2059**	2	28	7	6
Misc. (12)	3	7	**2**	0	0	0
Que. (141)	3	53	0	**85**	0	0
Rec. (20)	2	10	0	0	**4**	4
Req. (34)	1	16	0	1	2	**14**

We notice that most errors were made in sentences that belong to the class "Assertion" that were predicted as "Expression" (with 160 cases out of 1560 cases), and the class "Expression" which were predicted as "Assertion" (with 110 cases out of 2212 cases).

Intents that were misclassified to a high extent include: Miscellaneous, Question, Recommendation, and Request. They also tend to be classified as Expressions, with 7 out of 12 for Miscellaneous, 53 out of 141 for Question, 10 out of 20 for Recommendation, and 16 out of 34 for Request. We noticed that these four intents have a very low number of examples in total, which is the reason for the models having problems in classifying them correctly. Another plausible reason behind the faults in the Dialogue Acts prediction - beside the skewed dataset- is that sentences in the dataset are composite in nature, and thus encompassing diversified intentions in a single sentence. For example, in the sentence " التطرف والفساد وجهان لعملة واحدة، يتحركان في نفس الاتجاه وسلاح محاربتهما واحد إرساء دولة القانون", we can identify two Dialogue acts at the same time, Expression at the beginning and Recommendation in the final phrase. In Table 9, we present some examples, from ArSAS, where the model fails to predict the correct Dialogue acts.

Table 9. Examples of Labelling Errors on ArSAS

Orig. DA	Pred. DA	Sentence
Exp.	Ass.	حتى وكالة الانباء السعودية واس ومعظم المواقع السعودية حذفت خبر لقاء ولي العهد السعودي بقيادة اخوان اليمن بعد وقت قصير من نشره
Exp.	Que.	الى اهلنا في مصر، هل تعلمون أن جزيرتي تيران وصنافير هي حقا لكم ولاجدادكم ما قيمة ريالاتهم أمام حبة تراب لكم، هل هذا شعب الناصر؟
Ass.	Exp.	ولاد الناس انقرضوا للاسف وبعد ثورة ٢٥ يناير بانت كل الناس على حقيقتها
Rec.	Exp.	التطرف والفساد وجهان لعملة واحدة ،يتحركان في نفس الاتجاه وسلاح محاربتهما واحد إرساء دولة القانون
Req.	Exp.	أطالب ولي العهد السعودي بعد انتهاءه من اعتقال الأمراء أن يعتقل أمير الأحزان وأميرة بحجابي واميره بطبعي واميرة زوجي
Que.	Exp.	يعني اذا ما يؤيد حصار قطر صار خائن؟؟

We also compare our best model, AraBERTv0.2-Base, to the state-of-the-art on the same dataset performed by Algotiml et al. [23], who compared deep learning methods against machine learning methods. In their work, since the smallest two classes in the dataset "miscellaneous" and "recommendation" have only 60 and 109 tweets respectively, authors decided to merge these two classes into one and called it miscellaneous. Therefore, in order to compare with their

results, we merged the two classes, and fine-tuned our model again against the new dataset.

Table 10 shows a comparison between the best results of Algotiml et al. [23] in both machine learning and deep learning methods and our results using AraBERTv0.2-Base with/without doing resampling.

Table 10. Performance comparison with the state of the art on ArSAS

	Macro-F1	**Micro-F1**	**Accuracy**
SVM [23]	0.532	0.862	0.865
Bi-LSTM [23]	0.615	0.860	0.875
AraBERTv0.2-Base (-resampling)	0.53	**0.89**	**0.89**
AraBERTv0.2-Base (+resampling)	**0.62**	**0.89**	**0.89**

Our accuracy results outperform Algotiml et al. [23] results by 1.5%, and we notice that using resampling process, substantially raised the value of Macro-F1.

6.2 Model effectiveness on LevInt dataset

We applied the previous Arabert models on the Levantine Arabic dataset LevNet, which is balanced, and contains 873 sentences tagged with 8 speech acts. Table 11 presents the performance of the different models.

Table 11. Comparison between different BERT-Based models for LevInt speech act classification

Model	**Acc.**	**P.**	**R.**	**F1.**	**P.**	**R.**	**F1.**
		Macro Avg.			Weighted Avg.		
v0.1-orig	0.903	0.909	0.903	0.903	0.909	0.903	0.903
v1-orig	0.891	0.895	0.893	0.892	0.895	0.891	0.891
v0.2-base	**0.937**	**0.939**	**0.938**	**0.937**	**0.940**	**0.937**	**0.937**
v2-base	0.909	0.911	0.907	0.908	0.912	0.909	0.909
v0.2-large	0.823	0.828	0.823	0.822	0.827	0.823	0.821
v2-large	0.874	0.879	0.875	0.874	0.881	0.874	0.875

We remark that the best classification performance of all the models is obtained using the AraBERTv0.2-base pre-trained model, based on all metrics. This result confirms our previous conclusions produced based on the ArSAS dataset.

We also compare the result of our best model, AraBERTv0.2-Base, to the state-of-the-art model performed by Joukhadar et al. [25] on the same dataset. Authors in [25] compared different machine learning methods, and the only metric used for evaluating the different multi-labelling classifiers was the accuracy. Their results show that the SVM classifier outperforms the rest of the classifiers with an accuracy of 86%. When compared to our best accuracy (93.7%) stated in Table 11, we can conclude that our model outperforms their model by 7.7%.

7. CONCLUSIONS

In this paper, we have investigated different BERT-based Dialogue Act recognition models for Arabic language. For all models, we exploited the "ArSAS" Arabic corpus which has more than 21K tweets annotated by six different speech acts, and the "LevInt" Arabic corpus which has 873 sentences tagged with 8 speech acts. We implemented different Bert-Based models (AraBERTv0.1-original, AraBERTv1-original, AraBERTv0.2-base, AraBERTv2-base, AraBERTv0.2-large, AraBERTv2-large), and compared the proposed models on ArSAS, and LevInt. The best results were achieved using AraBERTv0.2-base model with an accuracy of 89% and 94% respectively.

We also compared our best model with the state-of-art-model, on each data, and our model outperformed the state-of-the-art model by 1.5% and 7.7% respectively.

Future directions for this work revolve around the research on possible training resources, since the current data proved to be somewhat effective, but also presented numerous drawbacks, such as the imbalanced Dialogue Acts and the lack of adequate coverage. The study of real Arabic human-human conversations, and the application of its features to man-machine interactions, can yield useful insights. The new larger dataset, with real life situations and speech act sequences, will permit considering the whole context of the sentence in predicting the speech act of each utterance. Building a Dialectal Arabic Morphological Analyser (or even a simple light stemmer), and using it in the pre-processing steps, will allow to extract important features, such as dialect negation tools, which are usually concatenated with the word itself, such as مارح/I will not in Levantine, or معرفش/I don't know in Egyptian, which will eventually improve the correct dialogue acts recognition.

REFERENCES

[1] Paladines, J., Ramirez, J. (2020). A systematic literature review of intelligent tutoring systems with dialogue in natural language. IEEE Access, 8: 164246-164267. https://doi.org/10.1109/ACCESS.2020.3021383

[2] Muhetaer, P., Ayifu, M., Dawa, I., Silamu, W. (2020). A multi-lingual and text-speech dialog support system for e-health. In Journal of Physics: Conference Series, 1549(5): 052029. https://doi.org/10.1088/1742-6596/1549/5/052029

[3] Hüning, H., Mechtenberg, L., Wang, S.W. (2021). Detecting argumentative discourse in online chat experiments. Working Paper. https://openarchive.tk.mta.hu/445/.

[4] Kale, M., Rastogi, A. (2020). Template guided text generation for task-oriented dialogue. arXiv preprint arXiv:2004.15006. https://arxiv.org/abs/2004.15006.

[5] Habash, N.Y. (2010). Introduction to Arabic natural language processing. Synthesis Lectures on Human Language Technologies, 3(1): 1-187. https://doi.org/10.2200/S00277ED1V01Y201008HLT010

[6] Kumar, H., Agarwal, A., Dasgupta, R., Joshi, S. (2018). Dialogue act sequence labeling using hierarchical encoder with CRF. arXiv:1709.04250 [cs.CL].

[7] Bothe, C., Weber, C., Magg, S., Wermter, S. (2018). A context-based approach for dialogue act recognition using simple recurrent neural networks. arXiv preprint arXiv:1805.06280. https://arxiv.org/abs/1805.06280.

[8] Lee, J.Y., Dernoncourt, F. (2016). Sequential short-text classification with recurrent and convolutional neural networks. arXiv preprint arXiv:1603.03827. https://arxiv.org/abs/1603.03827.

[9] Chen, Z., Yang, R., Zhao, Z., Cai, D., He, X. (2018). Dialogue act recognition via CRF-attentive structured

network. arXiv:1711.05568 [cs.CL].
[10] Wan, Y., Yan, W., Gao, J., Zhao, Z., Wu, J., Philip, S.Y. (2018). Improved dynamic memory network for dialogue act classification with adversarial training. In 2018 IEEE International Conference on Big Data (Big Data), pp. 841-850. arXiv:1811.05021.
[11] Dai, Z., Fu, J., Zhu, Q., Cui, H., Qi, Y. (2020). Local contextual attention with hierarchical structure for dialogue act recognition. arXiv preprint arXiv:2003.06044. https://arxiv.org/abs/2003.06044.
[12] Saha, T., Patra, A.P., Saha, S., Bhattacharyya, P. (2020). A transformer based approach for identification of tweet acts. In 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1-8. https://doi.org/10.1109/IJCNN48605.2020.9207484
[13] Saha, T., Saha, S., Bhattacharyya, P. (2019). Tweet act classification: A deep learning based classifier for recognizing speech acts in twitter. In 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, pp. 1-8. https://doi.org/10.1109/IJCNN.2019.8851805
[14] Saha, T., Jayashree, S.R., Saha, S., Bhattacharyya, P. (2020). BERT-caps: A transformer-based capsule network for tweet act classification. IEEE Transactions on Computational Social Systems, 7(5): 1168-1179. https://doi.org/10.1109/TCSS.2020.3014128
[15] Saha, T., Upadhyaya, A., Saha, S., Bhattacharyya, P. (2021). Towards sentiment and emotion aided multi-modal speech act classification in Twitter. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 5727-5737. http://dx.doi.org/10.18653/v1/2021.naacl-main.456
[16] Cerisara, C., Kral, P., Lenc, L. (2018). On the effects of using word2vec representations in neural networks for dialogue act recognition. Computer Speech & Language, 47: 175-193. https://doi.org/10.1016/j.csl.2017.07.009
[17] Jahanbakhsh-Nagadeh, Z., Feizi-Derakhshi, M.R., Sharifi, A. (2019). A speech act classifier for persian texts and its application in identifying rumors. arXiv preprint arXiv:1901.03904. https://arxiv.org/abs/1901.03904.
[18] Shala, L., Rus, V., Graesser, A.C. (2010). Automated speech act classification in Arabic. Subjetividad y Procesos Cognitivos, 14: 284-292.
[19] Sherkawi L., Ghneim N., AlDakkak O. (2017). Arabic speech act recognition using bootstrapped rule based system. International Journal on Computer and Communications Networks, Computational Intelligence and Data Analytics, 1(1).
[20] Sherkawi, L., Ghneim, N., Dakkak, O.A. (2018). Arabic speech act recognition techniques. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 17(3): 1-12. https://doi.org/10.1145/3170576
[21] Graja, M., Jaoua, M., Belguith, L.H. (2013). Discriminative framework for spoken Tunisian dialect understanding. In International Conference on Statistical Language and Speech Processing, pp. 102-110. https://doi.org/10.1007/978-3-642-39593-2_9
[22] Elmadany, A., Abdou, S., Gheith, M. (2018). Improving dialogue act classification for spontaneous Arabic speech and instant messages at utterance level. arXiv preprint arXiv:1806.00522. https://arxiv.org/abs/1806.00522.
[23] Algotiml, B., Elmadany, A., Magdy, W. (2019). Arabic Tweet-Act: Speech Act Recognition for Arabic Asynchronous Conversations. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, pp. 183-191. http://dx.doi.org/10.18653/v1/W19-4620
[24] Elmadany, A., Mubarak, H., Magdy, W. (2018). Arsas: An Arabic speech-act and sentiment corpus of tweets. OSACT.
[25] Joukhadar, A., Saghergy, H., Kweider, L., Ghneim, N. (2019). Arabic dialogue act recognition for textual chatbot systems. In Proceedings of The First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019-Short Papers, pp. 43-49. https://aclanthology.org/2019.nsurl-1.7.
[26] Antoun, W., Baly, F., Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. arXiv preprint arXiv:2003.00104. https://arxiv.org/abs/2003.00104.
[27] Wolf, T., Debut, L., Sanh, V., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771. https://arxiv.org/abs/1910.03771.