

# ArSentBERT: fine-tuned bidirectional encoder representations from transformers model for Arabic sentiment classification

Mohamed Fawzy Abdelfattah<sup>1</sup>, Mohamed Waleed Fakhr<sup>2</sup>, Mohamed Abo Rizka<sup>1</sup>

<sup>1</sup>Department of Computer Science, School of Computing and Information Technology, Arab Academy for Science Technology and Maritime Transport Cairo, Cairo, Egypt

<sup>2</sup>Department of Computer Engineering, School of Engineering and Technology, Arab Academy for Science Technology and Maritime Transport Cairo, Cairo, Egypt

## Article Info

### Article history:

Received Apr 4, 2022

Revised Jul 7, 2022

Accepted Sep 5, 2022

### Keywords:

Arabic BERT models  
Arabic sentiment classification  
Classification  
Tokenization  
Transformer

## ABSTRACT

Sentiment analysis in the Arabic language is challenging because of its linguistic complexity. Arabic is complex in words, paragraphs, and sentence structure. Moreover, most Arabic documents contain multiple dialects, writing alphabets, and styles (e.g., Franco-Arab). Nevertheless, fine-tuned bidirectional encoder representations from transformers (BERT) models can provide a reasonable prediction accuracy for Arabic sentiment classification tasks. This paper presents a fine-tuning approach for BERT models for classifying Arabic sentiments. It uses Arabic BERT pre-trained models and tokenizers and includes three stages. The first stage is text preprocessing and data cleaning. The second stage uses transfer-learning of the pre-trained models' weights and trains all encoder layers. The third stage uses a fully connected layer and a drop-out layer for classification. We tested our fine-tuned models on five different datasets that contain reviews in Arabic with different dialects and compared the results to 11 state-of-the-art models. The experiment results show that our models provide better prediction accuracy than our competitors. We show that the choice of the pre-trained BERT model and the tokenizer type improves the accuracy of Arabic sentiment classification.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Mohamed Fawzy Abdelfattah

Department of Computer Science, School of Computing and Information Technology

Arab Academy for Science Technology and Maritime Transport Cairo

Sheraton Al Matar, El Nozha, Cairo Governorate 4471344, Egypt

Email: mfawzy22@gmail.com

## 1. INTRODUCTION

With the rise of Arabic content shared by users through the internet, sentiment analysis has become essential to analyze people's thoughts and opinions in many applications, such as e-commerce, healthcare systems, and social networks. People in 26 countries across the Middle East and North Africa speak Arabic and use it to express opinions and thoughts on the internet and social media platforms [1]. However, Arabic has different dialects and writing styles. For example, Franco-Arabic consists of Arabic keywords written using Latin characters and numbers. In addition, the language has some limitations due to its complex orthography, morphology, and syntax. Therefore, sentiment analysis for the Arabic language is a challenging task [2]. Sentiment analysis (SA) is a natural language processing (NLP) research field that determines people's opinions, sentiments, and emotions. SA has three main classification types: sentence [3], document, and aspect [4]. SA methods are categorized into supervised, semi-supervised, and unsupervised machine learning approaches for sentiment classification. Most sentiment classification approaches fall into the

supervised category. The superior classical machine learning algorithms support vector machines (SVM), Naïve Bayes (NB) [5], and researchers have utilized them for Arabic sentiment classification [6]–[8]. Therefore, Arabic sentiment analysis has become a research interest for many researchers.

Recently deep learning has been used extensively in English language sentiment classification. Socher *et al.* [9] used the recurrent neural network (RNN) approach, which is trained on a sentiment tree bank. This approach improved the prediction accuracy of SA English text. However, there is less use of deep learning in Arabic SA than in English. Ghanem *et al.* [10] used a convolutional neural network (CNN) model for sentiment classification tasks and a stanford segmenter to perform tokenization and normalization of tweets. They also utilized word embedding Word2vec for the Arabic sentiment tweets dataset (ASTD). Alhumoduh *et al.* [11] used an long short-term memory CNN (LSTM-CNN) model with two classes (positive and negative) to classify tweets in the ASTD. Several recent studies [12], [13] have trained the bidirectional encoder representations from transformers (BERT) model on Wikipedia and Oscar datasets for the Arabic language. In addition to that, several recent studies [14], [15] have fine-tuned the Arabic BERT model [13] for downstream task SA. the drawback identified from the analysis of existing literature are: i) models not tested on different datasets; ii) some models ignore the context meaning of the sentence; iii) the model using context like BERT fined-tuned using general pre-trained models that affect models performance; and iv) there is room for improvement for reported prediction accuracy.

This work presents a fine-tuning technique for Arabic SA using Arabic pre-trained BERT models [12], [13], where one of them was chosen based on the dataset domain to maximize model performance and improve prediction accuracy. In order to achieve the best results, hyperparameters were chosen using population-based training [16]. This paper is structured as follows: section 2 proposes the research method, after which section 3 explains the experiments and results. Finally, section 4 introduces the conclusion and future work.

## 2. METHOD

The proposed approach comprises three stages: text preprocessing, fine-tuning the Arabic BERT model, and presenting the classification layer, as shown in Figure 1. The text preprocessing stage is responsible for data cleaning and text tokenization that feeds the model. The fine-tuning stage contains the pre-trained Arabic BERT model to initialize model weights for training. The classification stage contains a fully connected layer and a drop-out layer responsible for prediction.

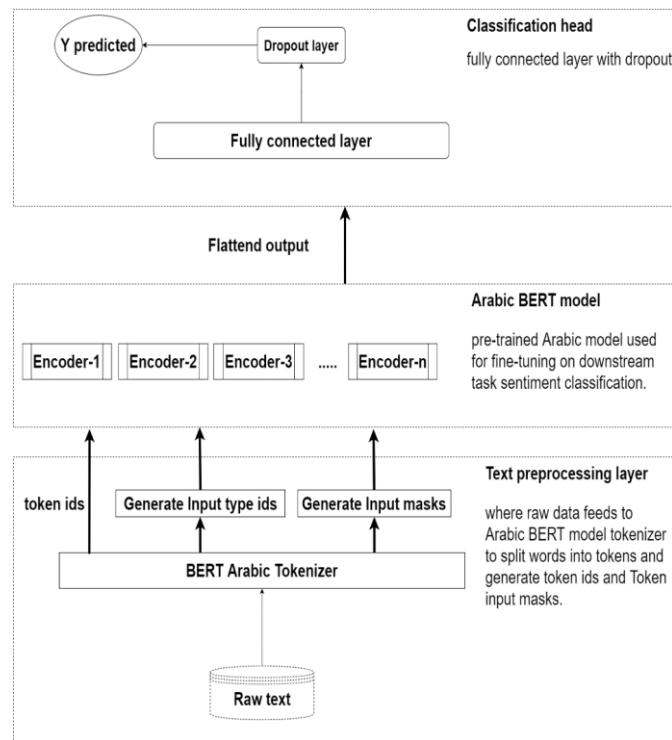


Figure 1. Proposed stages for fine-tuning

### 2.1. Text preprocessing

The first stage contains two steps: data cleaning and text tokenization. The first step is to remove the special characters, URLs, And hashtags when available; the second step is text tokenization, in which the sequence of strings can be split into a list of tokens based on the tokenizer type, such as character tokenizer, sub-word tokenizer, and word tokenizer. The original BERT tokenizer [17] was trained using WordPeice tokenization, in which a word can be split into multiple sub-words. Due to Arabic's morphological complexity, character and sub-word tokenizers are not a good fit for tokenization in Arabic. Therefore, we used two different pre-trained Arabic BERT tokenizers that use word-level tokenizers and applied a specific tokenizer on each dataset selected based on its context: Antoun *et al.* [13] tokenizer for social media datasets, such as Arabic–Jordanian tweets, and Safaya *et al.* [12] general-purpose tokenizer that works better on other datasets, such as Goodreads and booking.

### 2.2. Fine-tuning the Arabic bidirectional encoder representations from transformers model

In the second stage, the models were fine-tuned using Arabic BERT tokenizer and pre-trained Arabic BERT models [12], [13] for downstream task SA for Arabic language. In this approach, we used two different Arabic BERT models. We used one of the pre-trained models for each dataset we chose based on its context. We used Antoun *et al.* [13] pre-trained model to achieve better sentiment classification of social media datasets, such as Arabic Jordanian general tweets (AJGT); the model contains 12 encoders that are used for Arabic tweets. Furthermore, we used Safaya *et al.* [12] pre-trained model to achieve better accuracy of results regarding general-purpose datasets, such as Goodreads and Booking; the number of encoders in this model is 24. To achieve the best results, we used a hyperparameter search technique to find the best hyperparameters that can be used for each dataset using population-based training [16]. The BERT output layer contains a pooling operation in which the output is concatenated and flattened. The output then passes to the dense layer and activation function Tanh to calculate the probability for the label that is passed to the final stage, the classification layer. The fine-tuning process first initializes model weights using the pre-trained BERT models to transfer the statistical knowledge of the pre-trained language models trained on a large corpus. The models that use Safaya *et al.* [12] model as a pre-trained model fine-tuned all the 24 encoder layers and 340 M training parameters with a vocabulary size of 32 K. Moreover, fine-tuned models that use Antoun *et al.* [13] model fine-tuned all 12 encoder layers and 136 M training parameters with a vocabulary size of 64 K.

### 2.3. Classification layer

In the third stage, we present the classifier, also known as the classification head for the BERT model, as shown in Figure 2. It contains a drop-out layer for regularization and preventing overfitting and a linear layer to predict the output. The number of input features for the fully connected layer is different based on the pre-trained model architecture: Safaya *et al.* [12] used the BERT large model architecture, whereas Antoun *et al.* [13] used the medium architecture; correspondingly, for the BERT large architecture, the number of input features is 1,024, whereas for the BERT medium architecture, it is 768.

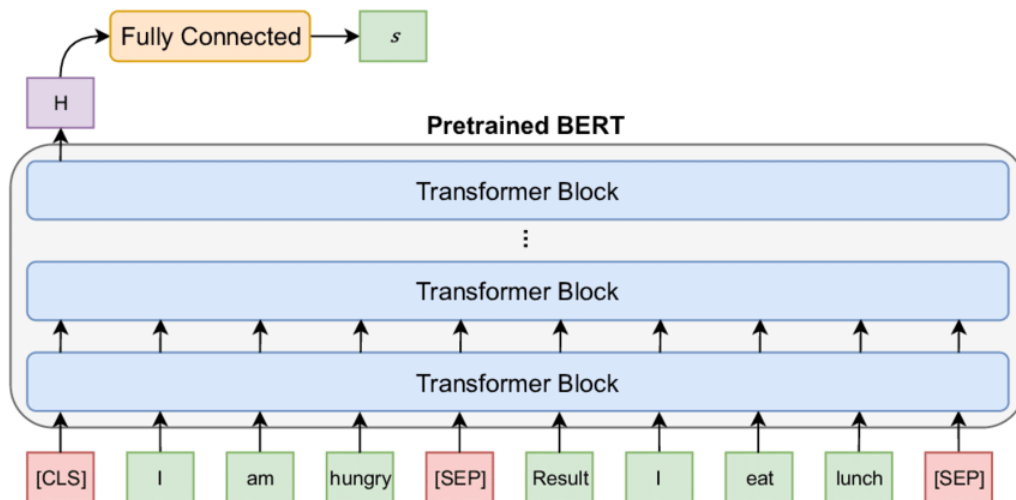


Figure 2. Proposed approach: the classification head

### 3. RESULTS AND DISCUSSION

Five datasets were used in this study, as described in Table 1, to benchmark the model results compared to those of other studies to evaluate our proposed approach. When available, the splits and text preprocessing provided by the dataset authors were used in this study. When they were unavailable, we split the dataset with consideration of the distribution of the dataset classes. Thus, we split the data into train, validation, and test classes, with a distribution of 70%, 10%, and 20%, respectively. In addition, we removed stopwords and hashtags from the Twitter datasets and URLs from the tweets for text preprocessing. Next, we fed these data in chunks to BERT tokenizers to tokenize the raw data and pass it to the model. We then applied different hyperparameters, as described in Table 2, to achieve best model performance, which is comparable with those in other studies. All experiments are based on Adam optimizer, a fine-tuned version of Adam optimizer.

Table 1. Datasets statistics

Dataset	Language	Samples	Classes
ASTD	MSA	1,000	2
LABR	DA	63,000	2
HARD	MSA-DA	93,700	2
AJGT	MSA-DA	1,800	2
ArSenTD-Lev	DA	4,000	3

Table 2. The fine-tuned Arabic BERT models' hyperparameters

Dataset	Batch_size	Drop-out	Max_length	Learning rate	Pre-trained model	Epochs
HARD	128	0.1	64	1e-5	bert-large-arabic	3
LABR	32	0.1	64	1e-5	bert-large-arabic	3
ArSenTD-Lev	512	0.1	128	5e-5	bert-base-twitter	50
AJGT	256	0.1	128	2e-5	bert-base-twitter	20
ASTD	64	0.2	140	5e-5	bert-large-arabic	20

#### 3.1. Datasets

The experiment was tested on five datasets. The datasets chosen in this experiment were collected from different data sources. It contains different dialects for Arabic speakers from different countries. All datasets' class distributions are balanced except goodreads review.

- ASTD: the ASTD [18] contains 10 K Arabic reviews collected from Twitter in 2015. It has different dialects. The tweets are labeled as “positive,” “negative,” “neutral,” and “mixed.” We used the balanced version of this dataset in the benchmark referred to as ASTD-B, which contains only balanced positive and negative examples.
- HARD: the hotel Arabic reviews dataset (HARD) [19] contains 93,700 reviews from 1,858 hotels contributed by 30,899 customers and collected in 2016. These reviews are annotated as “positive” and “negative.” We used the balanced dataset version; the unbalanced version had 373,750 reviews.
- AJGT: the AJGT [20] has 1,800 tweets annotated as “positive” and “negative.”
- ArSenTD-Lev: the Arabic sentiment twitter dataset for Levantine (ArSenTD-Lev) [21] contains 4,000 tweets written in the Levantine dialect with annotations for sentiment, topic, and sentiment target. There are five classes: “negative,” “neutral,” “positive,” “very negative,” and “very positive.” Our approach used three classes only: “positive,” “negative,” and “neutral.”
- LABR: the large-scale arabic book reviews (LABR) [7] contain 63,000 book reviews in Arabic. Reviews with ratings of 1 and 2 are considered negative, whereas reviews with ratings of 4 and 5 are considered positive. Reviews with a rating of 3 were discarded.

#### 3.2. Results and discussion

The proposed approach used two different pre-trained models. We chose one of the pre-trained models based on the dataset type to provide us with better-contextualized weights to initialize the model. Our model outperforms the state-of-the-art models AraBERT [13] and Choukhi *et al.* [14]. While both models use BERT architecture like the proposed approach, the main difference is that Choukhi *et al.* [14] uses BERT model medium architecture containing eight encoder layers without a text cleaning step. The proposed approach uses 12-encoder BERT architecture for Twitter and 24-encoder BERT architecture for general-purpose datasets, such as goodreads and hotel reviews. The data cleaning is applied before tokenization to remove the dataset noise in the text preprocessing step. Moreover, both approaches use one pre-trained model for all datasets, whereas we used two pre-trained models based on the dataset context. Our approach uses a

hyperparameter search to find the optimal parameters for each dataset using population-based training [16]. In addition, we performed some exploratory data analysis, as shown in Figure 3, to get the maximum length of words needed per dataset. Table 3 describes all the results of the models; in comparison to other studies, we reported the accuracy to compare with them because an evaluation metric, such as the F1-score, is not available; however, F1-score in the proposed approach is described in Table 4. All models' results were shared on GitHub.

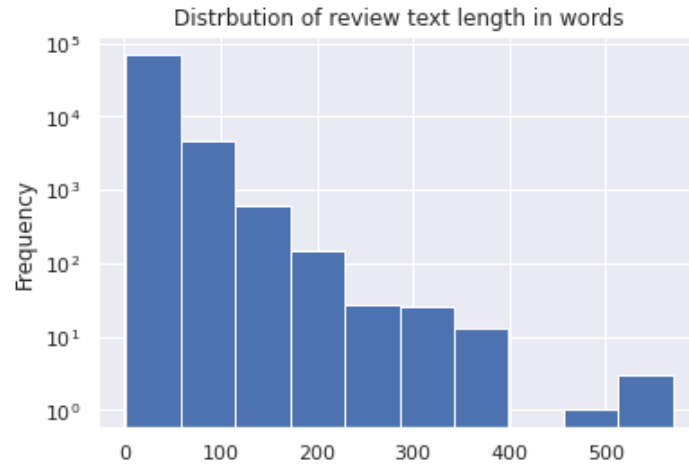


Figure 3. Exploratory data analysis for maximum length

Table 3. Comparison of existing models with our model

Approach	ASTD	LABR	AJGT	HARD	ArSenTD-Lev
CNN [10]	79	-	-	-	-
LSTM [22]	81	71	-	-	-
LSTM-CNN [11]	81	-	-	-	-
CNN-CROW [23]	72.14	-	-	-	-
DE-CNN-G1 [24]	82.48	-	93.06	-	-
LR [25]	87.10	84.97	-	-	-
GNB [25]	86	85	-	-	-
ARABERT-BASE [12]	71.4	-	-	-	55.2
hULMONA [26]	69.9	-	-	95.7	52.4
ARABERT [13]	92.6	86.7	93.8	96.2	59.4
Choukhi <i>et al.</i> [14]	91	87	96.11	95	75
The proposed approach	93.61	92.21	96.11	96.42	75.17

Table 4. The proposed approach F1-score

Dataset	Classes	Precision	Recall	F1-macro
ASTD	Negative	93.28	93.98	93.61
	Positive	93.94	93.58	
LABR	Negative	81.09	69.32	84.79
	Positive	93.99	96.89	
AJGT	Negative	96.42	97.78	96.11
	Positive	97.70	94.44	
HARD	Negative	97.33	95.47	96.42
	Positive	95.55	97.38	
ArSenTD-Lev	Negative	82.11	80.80	74.24
	Positive	71.74	78.75	
	Neutral	68.67	64.04	

#### 4. CONCLUSION





SA can result in reasonable accuracy if hyperparameters are optimized for the dataset, and there is room for improvement. More pre-trained BERT models, such as social media datasets and hate speech models, are needed in Arabic for specific tasks. This paper proposes an Arabic BERT model fine-tuned with regard to downstream task sentiment classification. The proposed approach achieves better accuracy than the current state-of-the-art models do since we used two different pre-trained Arabic BERT models. Each dataset

was trained by one pre-trained Arabic BERT model based on the dataset's context to help initialize the weights with better context values. Finally, according to our research, a BERT model with a few examples does not achieve significant improvement. It might need a data augmentation technique for oversampling the training dataset. Therefore, future studies that consider data augmentation techniques for low-data resources in Arabic are recommended.





## REFERENCES

- [1] S. Shorman and M. Al-Shoqran, "Analytical Study to Review of Arabic Language Learning Using Internet Websites," *International Journal of Computer Science and Information Technology*, vol. 11, no. 02, pp. 37–44, Apr. 2019, doi: 10.5121/ijcsit.2019.11204.
- [2] O. Oueslati, E. Cambria, M. Ben HajHmida, and H. Ounelli, "A review of sentiment analysis research in Arabic language," *Future Generation Computer Systems*, vol. 112, pp. 408–430, Nov. 2020, doi: 10.1016/j.future.2020.05.034.
- [3] N. Farra, E. Challita, R. A. Assi, and H. Hajj, "Sentence-Level and Document-Level Sentiment Mining for Arabic Texts," in *2010 IEEE International Conference on Data Mining Workshops*, Dec. 2010, pp. 1114–1119, doi: 10.1109/ICDMW.2010.95.
- [4] H. Zhou and F. Song, "Aspect-Level Sentiment Analysis Based on a Generalized Probabilistic Topic and Syntax Model," *The 28th Florida Artificial Intelligence Research Society Conference*, pp. 241–246.
- [5] A. Alsayat and N. Elmitwally, "A comprehensive study for Arabic Sentiment Analysis (Challenges and Applications)," *Egyptian Informatics Journal*, vol. 21, no. 1, pp. 7–12, Mar. 2020, doi: 10.1016/j.eij.2019.06.001.
- [6] H. AL-Rubaiee, R. Qiu, and D. Li, "The Importance of Neutral Class in Sentiment Analysis of Arabic Tweets," *International Journal of Computer Science and Information Technology*, vol. 8, no. 2, pp. 17–31, Apr. 2016, doi: 10.5121/ijcsit.2016.8202.
- [7] M. Aly and A. Atiya, "LABR: A Large Scale Arabic Book Reviews Dataset," *The 51st Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 494–498, doi: 10.13140/2.1.3960.5761.
- [8] H. ElSahar and S. R. El-Beltagy, "Building Large Arabic Multi-domain Resources for Sentiment Analysis," *Computational Linguistics and Intelligent Text Processing*, vol. 9042, pp. 23–34, Apr. 2015, doi: 978-3-319-18117-2\_2.
- [9] R. Socher *et al.*, "Recursive deep models for semantic compositionality over a sentiment treebank," *Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642.
- [10] B. Ghanem, J. Karoui, F. Benamara, V. Moriceau, and P. Rosso, "IDAT at FIRE2019," in *Proceedings of the 11th Forum for Information Retrieval Evaluation*, Dec. 2019, pp. 10–13, doi: 10.1145/3368567.3368585.
- [11] S. Alhumoud, T. Albuhaireh, and W. Alohaideb, "Hybrid Sentiment Analyser for Arabic Tweets using R," in *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 2015, pp. 417–424, doi: 10.5220/0005616204170424.
- [12] A. Safaya, M. Abdullatif, and D. Yuret, "KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 2020, pp. 2054–2059, doi: 10.18653/v1/2020.semeval-1.271.
- [13] W. Antoun, F. Baly, and H. Hajj, "ArABERT: Transformer-based Model for Arabic Language Understanding," *The 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pp. 9–15, doi: 10.48550/arXiv.2003.00104.
- [14] H. Chouikhi, H. Chniter, and F. Jarray, "Arabic Sentiment Analysis Using BERT Model," *Advances in Computational collective intelligence book*, 2021, pp. 621–632, doi: 10.1007/978-3-030-88113-9\_50.
- [15] H. EL Moubtahij, H. Abdelali, and E. B. Tazi, "ArABERT transformer model for Arabic comments and reviews analysis," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 1, p. 379, Mar. 2022, doi: 10.11591/ijai.v11.i1.pp379-387.
- [16] M. Jaderberg *et al.*, "Population Based Training of Neural Networks," *ArXiv preprint*, doi: 10.48550/arXiv.1711.09846.
- [17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
- [18] M. Nabil, M. Aly, and A. Atiya, "ASTD: Arabic Sentiment Tweets Dataset," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2515–2519, doi: 10.18653/v1/D15-1299.
- [19] A. Elnagar, Y. S. Khalifa, and A. Einea, "Hotel Arabic-Reviews Dataset Construction for Sentiment Analysis Applications," *Intelligent Natural Language Processing: Trends and Applications. Studies in Computational Intelligence*, Springer, Cham, vol. 740, pp. 35–52, 2018, doi: 10.1007/978-3-319-67056-0\_3.
- [20] K. M. Alomari, H. M. ElSherif, and K. Shaalan, "Arabic Tweets Sentimental Analysis Using Machine Learning," *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 2017, pp. 602–610.
- [21] R. Baly, A. Khaddaj, H. Hajj, W. El-Hajj, and K. Shaban, "ArSentD-LEV: A Multi-Topic Corpus for Target-based Sentiment Analysis in Arabic Levantine Tweets," *The 3rd workshop on open-source Arabic Corpora and processing tools of the 11th International Conference on Language Resources and Evaluation*, pp. 37–43, doi: 10.48550/arXiv.1906.01830.
- [22] A. Shoukry and A. Rafea, "Sentence-level Arabic sentiment analysis," in *2012 International Conference on Collaboration Technologies and Systems (CTS)*, May 2012, pp. 546–550, doi: 10.1109/CTS.2012.6261103.
- [23] R. Eskander and O. Rambow, "SLSA: A Sentiment Lexicon for Standard Arabic," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2545–2550, doi: 10.18653/v1/D15-1304.
- [24] A. Dahou, M. A. Elaziz, J. Zhou, and S. Xiong, "Arabic Sentiment Classification Using Convolutional Neural Network and Differential Evolution Algorithm," *Computational Intelligence and Neuroscience*, vol. 2019, pp. 1–16, Feb. 2019, doi: 10.1155/2019/2537689.
- [25] S. Harrat, K. Meftouh, and K. Smaili, "Machine translation for Arabic dialects (survey)," *Information Processing & Management*, vol. 56, no. 2, pp. 262–273, Mar. 2019, doi: 10.1016/j.ipm.2017.08.003.
- [26] O. ElJundi, W. Antoun, N. El Droubi, H. Hajj, W. El-Hajj, and K. Shaban, "hULMonA: The Universal Language Model in Arabic," in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, 2019, pp. 68–77, doi: 10.18653/v1/W19-4608.





**BIOGRAPHIES OF AUTHORS**

**Mohamed Fawzy Abdelfattah**     he is a master's student in Computer Science. His focus is on deep neural networks, precisely natural language processing, and its use for low-resource languages, such as Arabic. He is working as a senior software engineer and has over seven years of experience focusing on information retrieval, specifically search engines, documents retrieval, and ranking algorithms. His research interests include text classification, language generation, and text summarization. His first paper was published at the IEEE conference. He can be contacted via email: mfawzy22@gmail.com.



**Mohamed Waleed Fakhr**     finished his Ph.D at the University of Waterloo, Canada, in 1993, in the field of neural networks and machine learning. He then joined the speech research lab at NORTEL, Montreal, Canada, for five years, where he was a researcher investigating and implementing different speech processing, speech recognition, language modeling, and statistical error analysis techniques. He has two patents with NORTEL. Since 1999, he has been a professor with the Arab Academy for Science and Technology (Cairo, Egypt), with three years at the University of Bahrain. He has been performing research in the areas of time series forecasting, deep neural networks, natural language processing, and privacy-preserving computing. He can be contacted via email: waleedfakhr@yahoo.com.



**Mohamed Abo Rizka**     received his BS degree from MTC, Cairo, Egypt, in 1989; MS degree from MTC, Cairo, Egypt, in 1995; and Ph.D from the University of Alabama in Huntsville, Alabama, United States, in 2002. From 2002 to 2004, he was a lecturer at MTC. From 2004 to 2007, he was chairman of the e-commerce department at the Faculty of Management and Information Technology, Arab Academy for Science Technology and Maritime Transport (AASTMT), Cairo, Egypt. From 2007 to 2011, he was associate dean at the Faculty of Management and Information Technology, AASTMT. From 2011 to 2017, he was a dean at the Center of Excellence, AASTMT. From 2017 onward, he has been a dean at the College of Computing and Information Technology (AASTMT). His research interest includes big data analytics, intelligent systems, e-learning, bioinformatics, distributed systems, and cloud computing. He can be contacted via email: m.aborizka@aast.edu.