IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 15, 2021, accepted July 30, 2021, date of publication August 3, 2021, date of current version August 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3102087

Combining Context-Aware Embeddings and an Attentional Deep Learning Model for Arabic Affect Analysis on Twitter

HANANE ELFAIK AND EL HABIB NFAOUI, (Member, IEEE)
LISAC Laboratory, Department of Computer Science, Faculty of Sciences Dhar EL Mahraz (FSDM), Sidi Mohamed Ben Abdellah University, Fez 30003, Morocco

Corresponding author: Hanane Elfaik (hanane.elfaik@usmba.ac.ma)

ABSTRACT Affect analysis has recently attracted a great deal of attention due to the rapid development of online social platforms (i.e., Twitter, Facebook). Affect analysis is a part of a broader area of affective computing that aims to detect and grasp human emotions or affects within a piece of writing. Context awareness is very relevant for identifying human emotions and affects behind a piece of text. Capturing the context of a piece of text is often perceived as a challenge. In addition to the own unique features of tweets (shortness, noisiness, short length, etc.), the Arabic language is characterized by its agglutination and morphological richness. In this paper, we address the problem of Arabic affect detection (multilabel emotion classification) by combining the transformer-based model for Arabic language understanding AraBERT and an attention-based LSTM-BiLSTM deep model. AraBERT generates the contextualized embedding, and the attention-based LSTM-BiLSTM determines the label-emotion of tweets by extracting both past and future contexts considering temporal information flow in both directions. Additionally, the attention mechanism is applied to the output of LSTM-BiLSTM to emphasize different words. Our proposed approach was evaluated using the reference dataset of SemEval-2018 Task 1 (Affect in Tweets). The comprehensive results show that the proposed approach outperforms eight current state-of-the-art and baseline methods, and it achieves significant accuracy (53.82%) compared to 1st place in SemEval2018-Task1: (Affect in Tweets) competition. In addition, our proposed model outperforms the best recently reported model in the literature, with an enhancement of 2.62% in accuracy.

INDEX TERMS Emotion analysis, AraBERT, deep learning, multilabel emotion classification, attention mechanism, Arabic language.

I. INTRODUCTION

The authors in [1] define emotion recognition as the process of identifying human emotion by merely depending on personal skills and interpretation, by automating the process, or by a semi-automated approach. Automated emotion recognition (a.k.a. affect detection) aims at detecting human affective states such as happiness, sadness, and love from various modalities, including text, image, audio, and video. As a specific natural language processing (NLP) task, emotion detection from text has been a promising research topic over the years, and considerable efforts have been made to build a perfect automated system capable of detecting correct human emotions from text. It is considered a multilabel classification problem, i.e., more than one emotion can be conveyed in a piece of writing. Thus, it presents an additional challenge above binary or multiclass classification problems. Automatic text emotion detection can help analyse user attitudes, sentiments, and feelings from online textual data such as tweets, Facebook status, product reviews, comments, blogs, and news reports and might be applied to various fields, e.g., chatbots, e-learning systems, customer services, and mental health monitoring.

Twitter has become a popular platform for people to communicate and express emotions and feelings [2]–[5]. Therefore, Twitter provides a large amount of valuable data for text emotion analysis. However, conducting emotional analysis on tweets is a challenging task that has received considerable research attention due to some properties and characteristics related to tweets [1], [6], [7]. Indeed, the language

The associate editor coordinating the review of this manuscript and approving it for publication was Haiyong Zheng.

used on Twitter is ubiquitous, informal, and unstructured, as tweets often contain acronyms, spelling mistakes, abbreviations, and non-standard punctuation. In addition, tweets are short, noisy, and sometimes multilingual. Furthermore, tweets might represent sarcasm or use slang. This paper focuses on Arabic affect analysis. Arabic is considered the fifth most extensively spoken language in the world and is recognized as the official or native language for XXII countries [8], [9]. Arabic is both highly ambiguous and morphologically rich and has an enormous number of dialectal variants. Furthermore, compared to English, there are fewer freely available resources for Arabic emotion analysis. These difficulties have fueled broad research interest in Arabic emotion analysis [10]–[14], particularly in multilabel emotion classification.

Pre-trained word embeddings proved to increase classification performances in many NLP tasks, and they can be fine-tuned on a down-stream task or used as numerical features. Context-free embeddings [15], [16] generate a single global representation for each word within a corpus, ignoring their context. In contrast, context-aware embeddings (also called context-sensitive or contextualized embeddings) generate word vector representations that dynamically change with respect to the polysemy in the context in which the words appear [17]. Using context-aware embeddings for Arabic affect analysis can address many challenges. First, words are represented based on the context in which they appear. Hence, the emotion recognizer can deal with the semantics of words, rather than only shallow features. For example, the word ''ذهب'' will have different embedding vectors related to the following two sentences, as they have different meanings (''**gold**'' in the first sentence and ''**went**'' in the second sentence): **S1** = **"ذهب علي كثير"** (Ali has a lot of gold) **S2** = **"ذهب علي بعيدا"** (Ali went away). Second, the contextualized embedding represents words as numerical vectors, which makes it simpler to quantify and identify the emotion of tweet regarding the shared polysemy in context between them. Thus, the intuition behind our proposal in this paper is to exploit context-aware embedding for Arabic emotion analysis. In particular, we use the transformer-based model for Arabic language AraBERT [18] as the semantic contextual embeddings, and then we forward them to a deep learning model designed especially for multilabel emotion classification.

Deep learning can be defined as a subfield of machine learning that consists of multiple hidden layers that are designed for complex modelling and feature extraction [19], [20]. Deep learning has led to breakthroughs in many NLP applications, such as Arabic sentiment and affect analysis [21]–[29].

To the best of our knowledge, while there is a large body of literature on Arabic sentiment analysis [30], there are few research papers on Arabic affect analysis. To help advance the state-of-the-art performance in this affect detection task, we propose a hybrid approach combining AraBERT as a semantic contextual embedding with attention-based LSTM-BiLSTM as a multilabel emotion classification deep model.

The key contributions of our work can be highlighted as follows:

1) We have proposed an attention-based LSTM-BiLSTM deep model to determine the label-emotion of an input tweet. The results show that our model is more effective, and it has successfully achieved the highest accuracy on Arabic affect analysis multilabel classification compared to the current state-of-the-art methods with an enhancement of **2.62%.**
2) We have performed extensive experiments using different versions of AraBERT and BERT Multilingual. The results show that the contextual word representation produced by the pre-trained AraBERTv02-large-without Pre-Segmentation performs slightly better than the other representation. It significantly addresses language ambiguity, performs the deep relationships among Arabic words, and captures their polysemy in the context.
3) We have tested the effectiveness of our model on the only public and available benchmark dataset for Arabic multilabel emotion detection, namely, SemEval-2018 [31]. Our model achieves considerably better performance beating all best performing models in SemEval2018-Task1: (Affect in Tweets) competition, reaching an accuracy of 53.82%.
4) With this research, we have overviewed and discussed the current state-of-the-art methods for Arabic emotion analysis. In particular, we highlighted the main approaches, major contributions, emotion models, features, evaluation metrics, and results.

The rest of this paper is organized as follows. Section II presents recent related work on Arabic affect analysis methods. Section III presents preliminaries. Section IV describes our proposal. Section V provides the experimental study, the results obtained, and a discussion. Finally, Section VI concludes the paper and outlines future work.

II. RELATED WORK

In this section, we present a summary of the previous work in affect analysis. The available research on Arabic emotion analysis approaches can be grouped into lexicon-based, machine learning, deep neural networks, and hybrid approaches.

One of the earliest works on Arabic emotion detection was proposed by [32]. They developed a lexicon-based approach to determine emotions in Arabic children's stories based on six basic emotions of Ekman in addition to two other categories: the ''Neutral'' category, which does not convey any emotion, and the ''Mixed'' category, which conveys multiple emotions. This approach was applied at the word, sentence, and document levels to extract the emotions. After the preprocessing step, and using the cosine similarity, they compare the sentences with the basic six emotions. Refer-

ence [33] considered Ekman's basic emotions to automatically detect emotions in Arabic tweets for standards and the slang Egyptian dialect. They collected 1605 tweets, each annotated by an average of 15 human annotators. Five preprocessing techniques were discussed. Finally, the average accuracy of all emotions was 64.3%. However, the authors focus on a specific topic about the Egyptian revolution in 2011. Furthermore, the size of the dataset is small (1605 tweets).

The lack of available emotional resources for the Arabic language is a major issue. To circumnavigate this issue, English lexicons are often translated into the Arabic language. Reference [34] proposed a lexicon-based approach to extract and predict emotions in Arabic texts. Based on the eight emotions of Plutchik, they used an existing emotion lexicon called the NRC Emotion Lexicon (EmoLex) [35]. EmoLex was created for English and consisted of 14182 terms. Then, the latter was translated into 20 different languages, including Arabic. However, the lexicon was reduced to 4279 Arabic terms after removing the terms conveying no emotions and duplicates caused by the automatic translation. Finally, they evaluated the performance of their approach by using 39 text excerpts collected from different online resources, and the results achieved an accuracy of 89.7%. Another work that attempted to address the lack of resources in Arabic emotion analysis was [36]. The authors developed an automatic system to annotate the training data by means of their embedded emojis. The emotional Arabic dataset was collected from Twitter. Based on four emotion classes, namely, anger, disgust, joy, and sadness, they considered two classifiers: Support Vector Machine (SVM) and Multinomial Naïve Bayes (MNB). The results show that the automatic labelling approach employing SVM and MNB was a more accurate manual labelling approach, and they achieved a 72.26% F1-measure SVM-based model and 75.35% MNB-based model.

Reference [37] created a dataset of 10065 tweets for Arabic emotion detection. The dataset was split in a balanced way across eight labels: sadness, joy, anger, surprise, sympathy, love, and fear in addition to the ''no emotion''. After the preprocessing step and the feature extraction techniques, the experimental study was conducted using different classifiers. The best results were achieved using the Naïve Bayes (NB) algorithm with an accuracy of 68.12%.

The combination of a lexicon-based approach and a multi-criteria decision-making approach for Arabic emotion analysis has proven to be relevant, as shown by [38]. They considered Ekman's basic emotions without the surprise emotion. They utilized the dataset proposed by [39] (1552 tweets) to create a lexicon for each emotion. They built five lexicons validated by two human experts with a high inter-annotator agreement. Then, using the emotion scoring algorithm, each tweet was represented by a vector of five emotion scores. Finally, they used a conditioned plot (co-plot) to classify the tweet by generating a two-dimensional graphic analysis space. The importance of this approach is in its ability to handle tweets with multiple emotions (multilabel classification). Another method focusing on multilabel classification was [40]. Based on Ekman's basic emotions, the authors proposed a fine-grained approach in which a given tweet may have multiple emotions (multilabel), each with possibly different intensities (multitarget). They built and annotated a dataset of 11503 tweets. The dataset was annotated by two native Arabic speakers.

The study of a different view on the granularity in emotion detection was proposed in [41]. Based on six emotions, namely, happiness, surprise, anger, and sadness, in addition to sarcasm expression, they proposed a time emotional analysis system that contains four components, namely, the annotating tweets process, classification at tweet/expression levels, clustering on some aspects, and analysing over specific times the distributions of people's emotions, expressions, and aspects.

In one of the largest and most comprehensive efforts to address the problem of emotion analysis on Twitter, [31] organized SemEval-2018 Task 1 (Affect in Tweets). The task involved five subtasks: Emotion Intensity Regression EI-reg: (''Given a tweet and an emotion E, determine the intensity of E that best represents the mental state of the tweeter''), Emotion Intensity Ordinal Classification EI-oc: (''Given a tweet and an emotion E, classify the tweet into one of four ordinal classes of intensity of E that best represents the mental state of the tweeter''), Valence (sentiment) regression V-reg: (''Given a tweet, determine the intensity of sentiment or valence V that best represents the mental state of the tweeter''), Valence ordinal classification V-oc: (''Given a tweet, classify it into one of seven ordinal classes, corresponding to various levels of positive and negative sentiment intensity, that best represents the mental state of the tweeter''), and Emotion classification E-c: (''Given a tweet, classify it as 'neutral or no emotion' or as one, or more, of eleven given emotions that best represent the mental state of the tweeter'') in three languages (English, Spanish and Arabic). A total of 75 teams participated in this task. The datasets were annotated using Best Worse Scaling (BWS), and they were made available to the community. Reference[1] [42] participated in SemEval-2018-Task1. They developed a model with a dense network and an LSTM deep network to identify and predict the intensity of the emotions conveyed in tweets. A combination of word2vec and doc2vec embeddings and a set of psycholinguistic features (e.g., from AffectiveTweets Weka-package) was used as an input to their system. Then, they applied a fully connected neural network architecture to obtain the results. Another method was proposed as a team in SemEval-2018 Task 1 for affect analysis of Arabic tweets [43]. They participated in all 5 subtasks. Several preprocessing steps and several features were evaluated along with different classification and regression methods. In addition, they use SVC (Support Vector Classifier) with L1 and L2 used as penalties, RC (Ridge Classification), RF (Random Forest), and Ensemble. SVC with L1 performed best. The authors achieved 1st place in subtask 5, and 3rd place

[1] https://competitions.codalab.org/competitions/17751

TABLE 1. A summary of the Arabic emotion detection sorted on the newest date.

Authors\Year	Approach	Detection method	Emotion model	Features	Dataset	Evaluation metric and Results
Alswaidan el al., 2020 [21]	Hybrid approach	Deep feature-based (DF) model and Human engineered feature based (HEF) model	Plutchik's model + love, optimism, and pessimism emotions	Stylistic, lexical, syntactic, and semantic features.	SemEval-2018 dataset	Accuracy, - HEF 44.8% - DF 50.5% - Hybrid (HEF + DF) 51.20%
Tawalbehe el al., 2019 [47]	Machine learning approach	SVM, NB	Sad, joy, disgust, anger	TF-IDF		Accuracy, SVM 80.6%, NB 95%
Abdullah et al., 2018 [42]	Deep learning approach	Dense Network + LSTM	Plutchik's model	Word embedding Document embedding Psychological Linguistic features	SemEval-2018 dataset	Accuracy 44.6%
Badaro *et al.*, 2018 [43]	Machine learning approach	SVC, RC, RF	Plutchik's model	N-grams, lexicons, Word embedding, Fast-Text	SemEval-2018 dataset	Accuracy 48.9%
Mulki *et al.*, 2018 [44]			Plutchik's model	TF-IDF	SemEval-2018 dataset	Accuracy 46.5%
Abdullah, et al., 2018 [45]	Deep learning approach	CNN, LSTM		Word embedding Document embedding Semantic features	SemEval-2018 dataset	
Jabreel et al., 2018 [46]	Machine and deep learning approach	Ensemble XG boost regressor, deep learning N-Channels	Anger, fear, joy, sadness	Lexicon Features, Embedding Features	SemEval-2018 dataset	Pearson V-reg: ENG (82%), ARA (82%)
Al-Khatib et al., 2017 [37]	Machine learning approach	Naive Bayes	Sadness, anger, joy, surprise, love, sympathy, fear, no emotion"	N-grams, feature vector, BOW	Twitter (10,065 tweets)	10 CV, Accuracy = 68.12%.
Al-A'abed et al., 2016 [34]	Lexicon-based approach	Lexicon based	Anger, anticipation, joy, fear, sad, trust			Accuracy = 89.7%
Hussien *et al.*, 2016 [36]	Machine learning approach	SVM and Multinomial NB	Anger, disgust, joy and sadness.	BOW, TF-IDF	Twitter	F1 measure: SVM 72.26% MNB 75.34%
Sayed *et al.*, 2016 [41]	Machine learning approach	Conditional Random Fields and AdaBoost	Sadness, happiness, anger, surprise and sarcasm.	Word features Tweet features Structure features	Twitter (10177 tweets)	
Abd Al-Aziz et al., 2015 [38]	Lexicon-based and Multi-Criteria Decision-Making approach	Co-Plot	Happiness, sadness, fear, anger, and disgust		Twitter	2-D graphical representation
Rabie et al., 2014 [33]	Machine Learning approach	SVM and NB	Ekman's model		Twitter (1605 tweets)	Accuracy = 64.3%.
El Gohary *et al.*, 2013 [32]	Lexicon-based approach	Lexicon based	Six basic emotions of Ekman + Neutral and Mixed category	Word, sentence, and document level	Arabic children's stories	Accuracy = 65%.

in subtasks 1 and 3. In addition, [44] developed a multilabel classification system to detect the emotions embedded in Arabic, Spanish and English tweets. The binary relevance transformation strategy was employed, and TF-IDF was used to generate the tweets' features in SemEval-2018 Task 1. Additionally, [45] presented the SEDAT(Sentiment and Emotion Detection in Arabic Text) system using deep learning models to predict the intensity of emotions and sentiments conveyed in Arabic tweets. They used word embeddings, document embeddings, psycholinguistic features through the AffectiveTweets package, Deepmoji, and unsupervised sentiment neurons. Then, those vectors were fed into several deep neural network architectures, namely, feed-forward, CNN, and LSTM, on SemEval-2018 Task 1's datasets to obtain the predictions. In [46]. They proposed an emotion detection system that has been utilized in SemEval-2018 Task1 (Affect in Tweets). The authors combined two deep learning models (N-Stream and ConvNets) and XGBoost regressor based on a set of embeddings and lexicon-based features. The results of their system outperformed the other approaches in the valence intensity regression task and the valence ordinal classification task for the Arabic version. Additionally, the authors in [47] presented an emotion detection system across four label emotions: sadness, joy, disgust, and anger for Arabic. They used TF-IDF as features for two machine learning classifiers, NB and SVM. The results yield an accuracy of 80.6% by SVM and 95% by NB.

A multilabel classification was employed to detect emotions in Arabic tweets [21]. The authors proposed three models, namely, the ''Human engineered feature-based (HEF)'' model, ''Deep feature-based (DF)'' model, and ''Hybrid model'', based on both HEF and DF. The HEF model exploited a set of syntactic, semantic, and lexical human engineered features. The DF model exploited a combination of embedding layers: Emoji2vec, AraVec, GloVeEmb, and FastTextEmb. The results demonstrated that the hybrid (HEF + DF) model achieved an accuracy of 51.20% with an enhancement of 2.3% over the best performing model [43] in the SemEval2018-Task1 competition: (Affect in Tweets) [31]. Table 1 summarizes the relevant Arabic affect methods reviewed in this paper and sorted on the newest date.

III. PRELIMINARIES

This section presents the necessary background for understanding the remainder of this paper, including the problem definition, affect detection, word embedding representation, and deep learning for multilabel emotion classification used to implement our proposal.

A. PROBLEM DEFINITION

In this paper, we address the affect detection problem in Arabic tweets. A tweet may have multiple emotional states (for example *joy, love, optimism).* In this case, the emotion classification of tweets is framed as a multilabel classification problem.

Let $D = \{(x_i; y_i)\}_{i=1}^{N}$ denote the dataset, which contains N tweets with corresponding labels $y_i = \{0; 1\}^Q$ representing either the presence or absence of a label in the tweet x_i, where Q indicates the total number of labels.

Let $x_i = \{t_{i1}, t_{i2}, \ldots, t_{ip}, \ldots, t_{in}\}$ indicate the i^{th} tweet, with t_{ip} denoting the p^{th} token in the i^{th} tweet and n being the number of tokens in the tweet.

The multilabel tweet classification task aims to classify an instance into a set of labels. Therefore, the task requires training a classifier $f : x_i \longmapsto \tilde{y}_i$ to assign the most relevant labels to a tweet. Table 2 presents the description of notations used in the rest of this paper.

TABLE 2. Symbol denotation.

Notation	Description
D	Twitter multilabel dataset
N	The number of tweets in D
Q	Total number of labels
x_i	The input tweet
n	The number of tokens in the i^{th} tweet
t_{ip}	The p^{th} token in the i^{th} tweet
y_i	The set of true labels of the i^{th} tweet
$\tilde{y}_i$	The set of labels predicted by the classifier of the i^{th} tweet

B. AFFECT ANALYSIS

This subsection provides some details about the emotion detection problem, the application of emotion analysis and the emotion models.

1) EMOTION DETECTION PROBLEM

Detecting human emotions is a laborious task because of the ambiguity and versatility of human emotions. The same emotions might be expressed in multiple ways, and multiple emotions some of the time have the same expressions. Additionally, emotions might be dependent on gender, personality, ethnicity, location, culture, and numerous other social, psychological, and individual boundaries. The emotion detection task can be performed depending on various sources of information, such as speech [1], [48], [49], textual [50]–[53], or visually [54], [55].

Emotion analysis from text is a more complex task than sentiment analysis. Although these two terms are sometimes utilized synonymously, they differ in definition when utilized in computer science [56]. According to the Oxford Dictionary, 'emotion' is ''a strong feeling deriving from one's circumstances, mood, or relationships with others'', whereas 'sentiment' is ''a view or opinion that is held or expressed''. Additionally, Cambridge Dictionary defines 'emotion' as

''a strong feeling such as love or anger, or strong feelings in general'' and 'sentiment' as ''a thought, opinion, or idea based on a feeling about a situation, or a way of thinking about something''. Generally, 'sentiment' is defined as the effect of 'emotion' [57]. In other words, sentiment analysis extracts subjective information from a piece of text and identifies the polarity of an attitude of a person towards another person, event, thing, or task. However, emotion analysis focuses on extracting how a person feels about another person, event, or thing based on predefined emotion models [55].

2) APPLICATION OF EMOTION ANALYSIS FROM TEXT

Emotion analysis has various applications in every aspect of our daily life, including making efficient e-learning frameworks according to the emotion of students, improving human-computer interactions, monitoring the mental health of individuals, improving business strategies based on customer emotions, analysing public emotion on any national, international or political event, recognizing potential criminals from analysing the emotions of people after an attack or crime, improving the performance of chatbots and other automatic feedback frameworks.

Furthermore, social media activities gave rise to the immense shared people's feelings and emotions. Indeed, text as a source of information is still the most common form of communication on social media. People express their emotions through social media posts such as Facebook status, tweets, comments on their own or other people's posts, microblogs, and product reviews. Analysing these texts and identifying emotion from their words and semantics is a difficult challenge. In addition, emotion analysis from text has been a promising research topic over the years, and extensive efforts have attempted to build an automated system capable of identifying correct human emotions from text.

3) EMOTION MODELS

From a psychological perspective, human emotions can be assembled based on emotion type, emotion intensity, and numerous other parameters, which can be completely combined and acknowledged into emotion models. There are various theories about how to represent emotions [58]. However, the most important and frequently utilized in existing approaches are *categorical* and *dimensional*.

- *Categorical Emotion Models:* Present a set of categories of emotions that are discrete from each other. In this respect, we find Ekman's emotion model that contains six basic emotions are *anger, disgust, fear, happiness, sadness* and *surprise* [59].
- *Dimensional Emotion Models:* Present a few dimensions with some parameters and characterize emotions according to those dimensions. Each emotion occupies a location in this space [1]–[5]. The more representative emotion models of this approach are Russell [60] and Plutchik [61]. Figure 1 describes the eight basic emotions of Plutchik [61].

As we can check out in the related work section, categorical approaches are the most commonly used. Most of the computational approaches are based on the categorical emotion model, because of its simplicity. Nevertheless, categorical emotion models may not satisfactorily cover all emotions because emotion categories are restricted. This is a significant advantage of dimensional emotion models that are not correlated to a specific emotional state and can capture subtle emotion concepts that differ slightly. In addition, a dimensional emotion model provides a way to estimate and measure the similarity between affective states [40]. There was no better emotion model than the others. The two models have both benefits and drawbacks. The choice of an emotion model is based on the set of emotions that we want to analyse. Table 3. summarizes a few basic emotion models used in the literature.

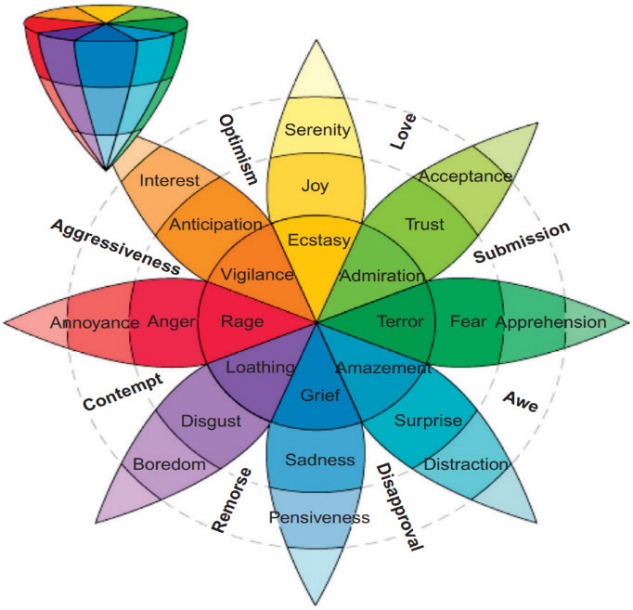

FIGURE 1. A set of eight basic emotions of Plutchik's emotion model [62]. Plutchik organizes these emotions on a wheel so that opposite emotions appear diametrically opposite to each other. Words closer to the centre have a higher intensity than those farther away.

C. WORD EMBEDDING REPRESENTATIONS

The word2vec model is the first meaningful representation for words created and developed by [63]. Since then, research has begun moving towards a variety of word2vec, such as GloVe [16] and fastText [64]. However, significant advances were accomplished with these models. They still needed and lacked contextualized information, which was handled by Bidirectional Encoder Representations from Transformers (BERT) [65]. BERT is a contextualized word representation model based on a multilayer bidirectional transformer encoder, where the transformer neural network utilizes parallel attention layers instead of sequential recurrence.

BERT is pre-trained on two unsupervised tasks: (i) a ''masked language model'' (Masked LM), where 15% of

TABLE 3. Emotion models (sorted on the emotion approach).

Model	Year	Basic emotions	Approach	Structure
Ekman [59]	1992	Anger, disgust, fear, joy, sadness, surprise.	Categorical	-
Shaver *et al.* [66]	1987	Anger, fear, joy, love, sadness, surprise.	Categorical	Tree
Oatley et al. [67]	1987	Anger, anxiety, disgust, happiness, sadness.	Categorical	-
Plutchik [61]	1980	Acceptance, admiration, aggressiveness, amazement, anger, annoyance, anticipation, apprehension, awe, boredom, contempt, disapproval, disgust, distraction, ecstasy, fear, grief, interest, joy, loathing, love, optimism, pensiveness, rage, remorse, sadness, serenity, submission, surprise, terror, trust, vigilance.	Dimensional	Wheel
Russell [60]	1980	Afraid, alarmed, angry, annoyed, aroused, astonished, at ease, bored, calm, content, delighted, depressed, distressed, droopy, excited, frustrated, glad, gloomy, happy, miserable, pleased, relaxed, sad, satisfied, serene, sleepy, tense, tired.	Dimensional	Valence, arousal
Oatley et al. [68]	1988	Admiration, anger, appreciation, disappointment, disliking, fear, fears-confirmed, gloating, gratification, gratitude, happy-for, hope, liking, pity, pride, sorry-for, relief, remorse, reproach, resentment, self-reproach, shame.	Dimensional	Tree
Lövheim [69]	2012	Anger/rage, contempt/disgust, distress/anguish, enjoyment/joy, fear/terror, interest/excitement, shame/humiliation, surprise/startle.	Dimensional	Cube

the tokens are randomly masked and replaced with the ''[MASK]'' token, then the model is trained to predict the masked tokens, and (ii) a ''Next Sentence Prediction'' (NSP) task, where the model is given a pair of sentences and is trained to predict and identify when the second one follows the first. BERT was trained on the BooksCorpus dataset (800 M words) [70] and text passages of English Wikipedia. There are two available pre-trained model sizes for BERT: BERT-Base and BERT-Large. Table 4 presents the specifications of the BERT-Base and BERT-Large models. The pre-trained publicly available BERT model and code for fine-tuning on a specific task are available online.[2,3]

Furthermore, many language-specific versions of BERT are available, which are trained on specific language text, including the following:

- *Multilingual BERT* [65] is pre-trained in the same way as monolingual BERT except using Wikipedia text from the top 100+ languages (*Arabic, Dutch, German, Spanish,...*). To account for the differences in the size of Wikipedia, using exponential smoothing, some languages are sub-sampled, and some are super-sampled.
- *AraBERT* [18] is the pre-trained BERT specifically for the Arabic language. It was trained on ∼70 M sentences or ∼23 GB of Arabic text with ∼3B words. The training corpora are a collection of publicly available large scale Arabic text (code and pre-trained models are publicly available). Figure 2 describes the model structure of AraBERT. Two pre-trained versions for AraBERT are available: AraBERT-v1 and AraBERT-v2 (base and large) with better vocabulary, more data, and more training. Table 4 presents the specifications of the AraBERT-v1 and AraBERT-v2 models.

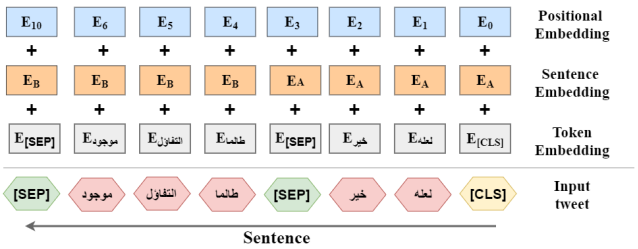

FIGURE 2. Model structure of AraBERT. Taking a tweet of two parts as an example, the input tweet is embedded by token embedding, sentence embedding and positional embedding from bottom to top. Using the AraBERT encoder, each part of the tweet is encoded into a vector, and using a transformer decoder with an activation layer, the score of each part is calculated.

D. LANGUAGE-SPECIFIC BERT FINE-TUNING

The pre-trained language model BERT can be fine-tuned to a specific task. Using a small corpus of task-specific data, fine-tuning BERT consists of adjusting pre-trained BERT model parameters to a specific task. For the purpose of the multilabel emotion classification task, a neural network layer is utilized on top of the fine-tuned BERT model. Indeed, the weights of the neural network and the weights of the BERT model are trained and fine-tuned correspondingly using task-specific data. Figure 3 illustrates an overall scheme of BERT fine-tuned for an affect analysis specific task.

E. BILSTM WITH ATTENTION LAYER

LSTM is an artificial recurrent neural network (RNN) architecture, which is constructed to deal with sequential data [71].

[2]https://github.com/google-research/bert
[3]https://github.com/huggingface/pytorch-transformers

TABLE 4. Specifications of BERT, BERT- multilingual, and AraBERT models.

	Number of transformer layers	Number of hidden units in each layer	Number of attention heads per hidden unit	Number of total parameters
BERT-Base	12	768	12	110,000,000
BERT-Large	24	1024	16	340,000,000
BERT-Base, Multilingual Cased	12	768	12	110,000,000
BERT-Base, Multilingual Uncased	12	768	12	110,000,000
AraBERT v1-Base	12	768	12	136,000,000
AraBERT v2-Base	12	768	12	136,000,000
AraBERT v2-Large	24	1024	16	371,000,000

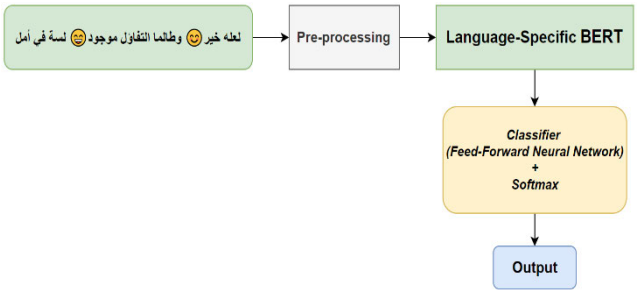

FIGURE 3. Language-specific BERT fine-tuned for affect analysis specific task.

In addition, LSTM captures long-term dependencies and addresses the problem of vanishing using its gates to manage the error gradient. The hidden state of an LSTM unit is computed by [71]

$$f_t = \sigma(W_f.x_t + U_f.h_{t-1} + b_f) \quad (1)$$
$$i_t = \sigma(W_i.x_t + U_i.h_{t-1} + b_i) \quad (2)$$
$$o_t = \sigma(W_o.x_t + U_o.h_{t-1} + b_o) \quad (3)$$
$$c_t = f_t \circ c_{t-1} + i_t \circ \tanh(W_c.x_t + U_c.h_{t-1} + b_c) \quad (4)$$
$$h_t = o_t \circ \tanh(c_t) \quad (5)$$

where x_t is the input at time t (the current word embeddings of a word in the tweet in the LSTM we worked with), f_t, i_t, o_t and c_t denote respectively the forget gate, the input gate, the output gate, and the memory cell, W_f, U_f, b_f are respectively two weights matrices and a bias vector for forget gate f. The denotation is similar to input gate i and output gate o, σ is the Softmax function, and $\circ$ is the Hadamard product. These gate units significantly help the LSTM model to remember information over multiple time steps [72].

In order to capture information from both directions. A Bidirectional LSTM (BiLSTM) makes two LSTMs, one takes the input in a forward direction and the other in a backward direction. Two hidden states $h_t^{forward}$ and $h_t^{backward}$ from these LSTM units are concatenated into a final hidden state h_t^{bilstm} [73]:

$$h_t^{bilstm} = h_t^{forward} \oplus h_t^{backward} \quad (6)$$

where $\oplus$ is the concatenation operator. Therefore, in order to enforce the contribution of essential words, we adopt the attention mechanism. The latter assigns a weight a_i to each token by means of a softmax function. The representation R, which is a weighted sum of all tokens, is then calculated as [22]:

$$R = \sum_{i=1}^{n} a_i h_i \quad (7)$$

where $a_i = \frac{exp(e_i)}{\sum_{j=1}^{n} exp(e_j)}$, $\sum_{i=1}^{n} a_i = 1$, $e_i = tanh(W_h.h_i + b_h)$, W_h and b_h are learned parameters, and h_i is the concatenation of the representations of the forward and backward LSTM. Then, we use the representation R produced by the attention layer to a fully connected layer in order to obtain the class probability distribution.

IV. OUR PROPOSAL

In this section, we propose a multilabel emotion classification model for tweets. Given a tweet, classify it as 'neutral or no emotion' or as one, or more, of eleven given emotions (*anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust*) that best represent the mental state of the tweeter. Figure 4 illustrates its overall architecture, which contains three major components: ''Tweet Preprocessing and Cleaning'', ''Mapping tweets to contextualized embeddings'', and ''Affect Classification''. The following subsections deeply describe each component.

A. TWEET PREPROCESSING AND CLEANING

Tweet preprocessing, which is the first step in our proposed method, converts Arabic tweets to a form that is appropriate and suitable for the multilabel emotion classification system. These preprocessing tasks included removing punctuation, Latin characters, stop words, diacritics, and digits, and investigating the tokenization process, normalization, and light stemming, Additionally, we enriched the tweets by transcribing their embedded emoji in its corresponding Arabic words. These linguistics are utilized to reduce the ambiguity and noisiness of the tweets to increase the accuracy and effectiveness of our proposal. In Table 5, we present the preprocessing

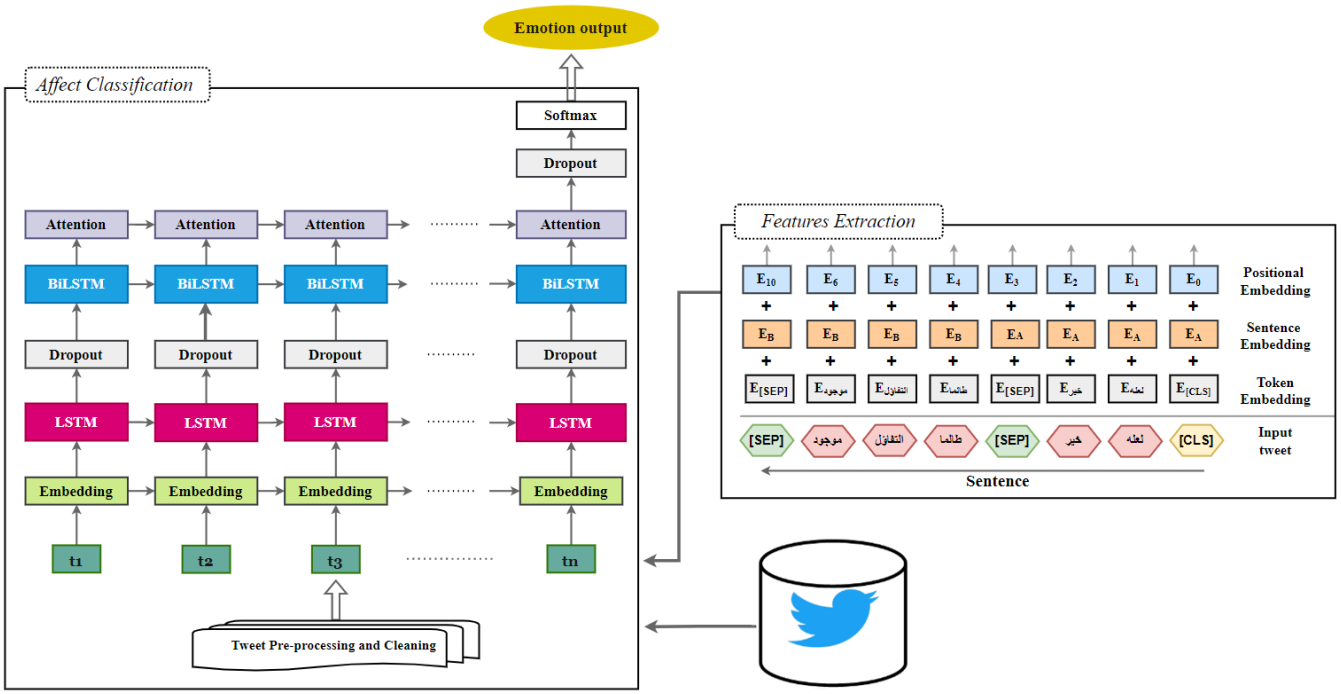

FIGURE 4. Overview of our Arabic affect recognition method.

techniques used and show how to apply them on a given example:

''لعله خير و طالما التفاؤل موجود لسة في أمل''

B. MAPPING TWEET TO CONTEXTUALIZED EMBEDDINGS

In this step, every token is mapped to an n-dimensional vector of real numbers. The emotion recognizer utilizes the language-specific BERT language model to map each token to the corresponding contextualized embedding. We used the BERT-Base Multilingual Cased, BERT-Base Multilingual Uncased models [65], and the AraBERT model [18], which was derived by further pretraining the original BERT-Base model on Arabic corpora.

The tokens are used as the input of the feature extraction step, and the output is the contextualized embeddings produced by different layers of BERT. Every token is represented as an n-dimensional vector that captures the context in which the token appears. In addition, the performance of our emotion recognizer was evaluated separately using different versions of BERT for the Arabic language.

C. AFFECT CLASSIFICATION

This section explains our affect classifier architecture based on the attentional LSTM-BiLSTM deep model. Considering the tweet as a sequence of words, LSTM has the advantage of recalling long-term special and temporal dependencies by connecting previous contexts to present contexts. Then, we added a BiLSTM layer to the LSTM layer to extract both past and future contexts by means of considering temporal information flow in both directions. Our system uses AraBERT pre-trained word embeddings to represent each token in the tweet by its corresponding contextualized embeddings. Then, we feed the obtained embeddings into our affect classifier to predict the corresponding overall emotions. We used the attention mechanism to emphasize different words and capture the most significant part of a target sentence. Notably, the attentionally weighted representation and the last hidden state are combined to obtain the final sentential representation. This trick can thereby improve the performance of multilabel classification.

V. EXPERIMENTS, RESULTS, AND DISCUSSION

This section evaluates the effectiveness of our proposed method using a benchmark multilabel SemEval2018-Ar dataset. Section V-A presents the dataset. Section V-B describes the evaluation metrics. In Section V-C, we introduce the state-of-the-art methods we have compared our system with. Section V-D presents the implementation details. Section V-E details the parameters setting. Finally, Sections V-F and V-G present and discuss the experimental results, respectively.

A. DATASET

In this paper, the experiments are conducted using the reference emotion detection SemEval-2018 (Affect in Tweets) dataset [31]. We used only the E-c (an emotion classification task) dataset for our experiment. To the best of our knowledge, this dataset is the only public and available benchmark dataset created for multilabel emotion detection in Arabic tweets. Each tweet is labelled as 'neutral or no emotion' or as

TABLE 5. Arabic text preprocessing techniques.

Preprocessing technique	Description	Example
Tokenization	Tokenization is a method for dividing texts into tokens.	لعله, خير,😂, و, طالما, التفاؤل, موجود,😁, لسة, في، أمل
Stop Word Removal	Remove the Stop words involves the elimination of insignificant words from tweets that do not have any meaning or indications about the content. Examples of these insignificant words are articles, conjunctions, pronouns (such as he/هو, she/هي, and they/هم), prepositions (such as from/من, to/الى, in/فى, and about/حول), and demonstratives, (such as this/هذا, these/هؤلاء, and there/اولئك). In addition, the Arabic circumstantial nouns indicate time and place (such as after/بعد, above/فوق, and beside/بجانب), signal words (such as first/اولا, second/ثانيا, and third/ثالثا).	لعله, خير,😂, طالما, التفاؤل, موجود,😁, لسة، أمل
Punctuation Removal	Remove the punctuations symbols such as {#, -, .,.,,,;,:,‹}, these symbols are not useful and insignificant in our approach.	لعله خير😂 طالما التفاؤل موجود 😁 لسة أمل
Latin Characters and Digits Removal	We remove the Latin characters and digits, which appear in the tweets because these are not helpful to detect the emotion conveyed in a tweet.	∅
Word Normalization	The normalization method aims to normalize certain letters that have different forms in the same word to one form. For example, the normalization of "آ" (aleph mad), "أ" (aleph with hamza on top), and "إ" (aleph with hamza at the bottom) to "ا" (aleph), the normalization of the letter "ى" to "ي", the normalization of "ؤ" (hamza on waw) and "ئ" (hamza on ya) to "ء" (hamza), and the letter "ة" to "ه".	لعله خير😂 طالما التفاءل موجود 😁 لسه امل
Diacritics Removal	The diacritics {ً,ٌ,ٍ,َ,ُ,ِ,ْ} are not used to extract and identify the Arabic roots. Furthermore, the letters that include the symbol "ّ" الشدة are duplicate.	∅
Emoji Replacement	A lexicon was created to contain the most frequent emojis in tweets. Additionally, we transcribed each emoji in the tweet to its corresponding Arabic word. The lexicon consisted of 100 emojis.	لعله خير خجل طالما التفاءل موجود ضحك لسه امل
Light Stemming	Light stemming is the process of reducing inflected words to their root/stem by stripping off a small set of prefixes and/or suffixes. For example: (تستلزم-لزم), (مكتب-كتب). Information Science Research Institute's (ISRI) stemmer developed by Taghva et al., 2005 was used in our proposed approach.	لعل خير خجل طال فاعل وجد ضحك لسه امل

one, or more, of eleven emotions (*anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust*). This dataset contains a total of 4381 tweets, 2278 in the training set, 585 in the development set, and 1518 in the test set. All these tweets are in Arabic. The statistics are shown in Table 6. Furthermore, Figure 5 shows the emotion correlations in SemEval-2018-Ar; orange indicates that two emotions are positively correlated, e.g., *joy* and *love*, whereas blue indicates that two emotions are negatively correlated, e.g., *anger* and *optimism*.

TABLE 6. The statistics of SemEval2018-Ar dataset.

N°	Emotion label	Number of tweets			
		Train	Dev	Test	Total
0	Anger	899	215	609	1723
1	Anticipation	209	57	158	421
2	Disgust	433	106	316	855
3	Fear	391	94	295	780
4	Joy	605	179	393	1177
5	Love	562	175	367	1104
6	Optimism	561	169	344	1074
7	Pessimism	499	125	377	1001
8	Sadness	842	217	579	1638
9	Surprise	47	13	38	98
10	Trust	120	36	77	233
	Total	**2278**	**585**	**1518**	**4381**

B. EVALUATION METRICS

This section presents the measures utilized to evaluate the performance of our emotion detection system. For the definitions below, y_i denotes the set of true labels of example x_i, $f(x_i) = \tilde{y}_i$ denotes the set of labels predicted by the classifier for the same examples, N is the number of examples, and Q is the total number of labels. All definitions refer to the multilabel classification setting utilized by the organizers of SemEval2018 Task 1 for the E-c (emotion classification) task. In addition, the evaluation metrics can be divided into two sub measures: *example-based measures* and *label-based measures* [74], [75].

1) EXAMPLE-BASED MEASURES

Accuracy for a single input example x_i is defined by the Jaccard similarity coefficient between the predicted label

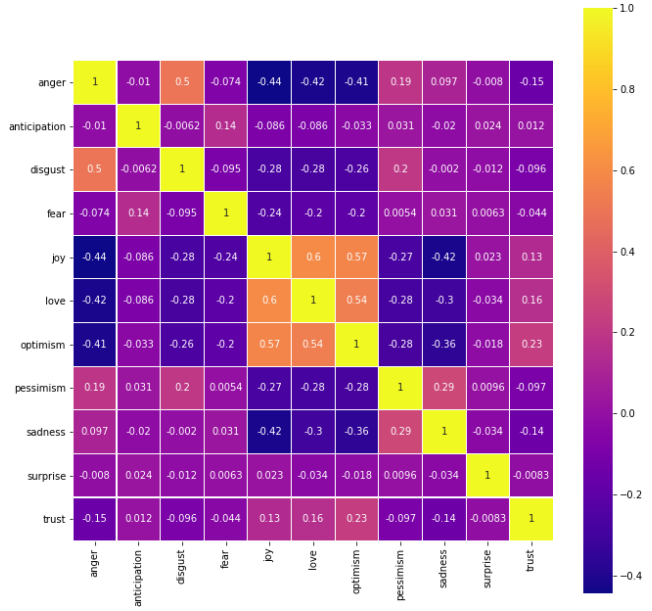

FIGURE 5. The emotion correlations in the SemEval2018-Ar dataset.

sets $\tilde{y}_i$ and true label sets y_i. Accuracy is defined as the micro-averaged across all examples in the dataset:

$$Accuracy\,(f) = \frac{1}{N}\sum_{i=1}^{N}\frac{|f(x_i)\cap \boldsymbol{y_i}|}{|f(x_i)\cup \boldsymbol{y_i}|} \tag{8}$$

Precision is defined as:

$$Precision\,(f) = \frac{1}{N}\sum_{i=1}^{N}\frac{|f(x_i)\cap \boldsymbol{y_i}|}{|\boldsymbol{y_i}|} \tag{9}$$

Recall is defined as:

$$Recall\,(f) = \frac{1}{N}\sum_{i=1}^{N}\frac{|f(x_i)\cap \boldsymbol{y_i}|}{|f(x_i)|} \tag{10}$$

F1- score is defined as the harmonic mean between precision and recall:

$$\begin{aligned} F1-score &= \frac{2\times Precision\,(f)\times Recall\,(f)}{Precision\,(f)+Recall\,(f)} \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{2\times|f(x_i)\cap \boldsymbol{y_i}|}{|f(x_i)|+|\boldsymbol{y_i}|} \end{aligned} \tag{11}$$

2) LABEL-BASED MEASURES

Macro-precision is defined as the precision averaged across all labels:

$$Macro-Precision = \frac{1}{Q}\sum_{j=1}^{Q}\frac{TP_j}{TP_j+FP_j} \tag{12}$$

Macro-recall is defined as the recall averaged across all labels:

$$Macro-Recall = \frac{1}{Q}\sum_{j=1}^{Q}\frac{TP_j}{TP_j+FN_j} \tag{13}$$

Macro-F1 is defined as the harmonic mean between precision and recall, where the average is calculated per label and then averaged across all labels. If p_j and r_j are the precision and recall for all $\lambda_j \in f(x_i)$ from $\lambda_j \in y_i$, the macro-F1 is defined as:

$$Macro-F1 = \frac{1}{Q}\sum_{j=1}^{Q}\frac{2\times p_j\times r_j}{p_j+r_j} \tag{14}$$

Micro-precision is defined as the precision averaged over all the example/label pairs:

$$Micro-Precision = \frac{\sum_{j=1}^{Q}TP_j}{\sum_{j=1}^{Q}TP_j+\sum_{j=1}^{Q}FP_j} \tag{15}$$

Micro-recall is defined as the recall averaged over all the example/label pairs:

$$Micro-Recall = \frac{\sum_{j=1}^{Q}TP_j}{\sum_{j=1}^{Q}TP_j+\sum_{j=1}^{Q}FN_j} \tag{16}$$

Micro-F1 is defined as the harmonic mean between micro-precision and micro-recall:

$$Micro-F1 = \frac{2\times Micro-Precision\times Micro-Recall}{Micro-Precision+Micro-Recall} \tag{17}$$

where TP_j, FP_j and FN_j are the number of true positives, false positives and false negative for the label λ_j considered as a binary class.

Additionally, we calculate the Area Under the Curve (AUC) [76], We intended to utilize AUC in order to plot the performance of the model across all labels. AUC values arrange from 0.0 to 1.0, with 0.5 being no more excellent and 1.0 is the ideal fit.

C. COMPARISON METHODS

We have compared our proposal with the baseline and state-of-the-art emotion analysis methods on the SemEval-2018-Ar dataset, including:

- *SVM-Unigrams [31]:* A baseline support vector machine system on the SemEval2018-Ar competition, trained using word unigrams as features.
- *Random Baseline [31]:* A baseline method on SemEval2018-Ar. It is a system that randomly guesses the prediction.
- *EMA [43]:* Achieved 1st place in subtask 5 (E-c). It utilizes various preprocessing steps (e.g., diacritics removal, normalization, emoji transcription and stemming). Additionally, it uses SVC (Support Vector Classifier) with L1 and L2 as penalties, RC (Ridge Classification), RF (Random Forest) and Ensemble. SVC with L1 performed best.
- *PARTNA [31]:* This method identifies the emotion of tweets using traditional machine learning approaches; therefore, this method uses the stemmer designed for handling tweets.

- *Tw-StAR [44]:* It develops a multilabel emotion classification system to detect the emotions embedded in Arabic, Spanish and English tweets. The binary relevance transformation strategy was employed, and TF-IDF was used to generate the tweets' features.
- *MEDIANteam [31]:* The system was submitted by the fifth-place winner team of the SemEval-2018 Task1: E-c challenge Arabic Ranking.
- *TeamUNCC [42]:* The main input to its system is a combination of word2vec and doc2vec embeddings and a set of psycholinguistic features (e.g., from AffectiveTweets Weka-package). It applies a fully connected neural network architecture to obtain the results.
- *Alswaidan and Menai [21]:* Proposed three models, namely, the ''Human engineered feature-based (HEF)'' model, ''Deep feature-based (DF)'' model, and ''Hybrid model'', based on both HEF and DF. The HEF model exploited a set of syntactic, semantic, and lexical human engineered features. The DF model exploited a combination of embedding layers: Emoji2vec, AraVec, GloVeEmb, and FastTextEmb.

D. IMPLEMENTATION DETAILS

The experiments were conducted on Colab Pro[4] with access to high memory VMs, Python 3.7.6 64 bits, Pytorch, Transformers, Python ML Libraries, and GPU- accelerated environment. For the implementation of the attentional LSTM-BiLTM deep model, we used the Keras[5] deep learning package. For fine-tuning BERT and AraBERT, we use the Huggingface[6] transformer library.

E. PARAMETER SETTING

We carry out a set of parameterization experiments to find those settings that obtain the best results. For this purpose, the number of epochs is 10 for all the experiments. To avoid the overfitting problem and to ensure the effectiveness of our method, we employ the dropout layer, with a rate of 0.25, and we also adopt the L2 regularization technique to reduce the size of large weights. In addition, we utilize the binary cross-entropy loss function to train our model, and we use the rectified linear unit (ReLu) with a batch size of 4. To classify the representation obtained from the final layer, Softmax was utilized. Furthermore, we use the RMSProp optimizer [77] to tune the learning rate. These parameters are given in Table 7. Additionally, for BERT fine-tuning, we used a maximum sequence of 200, batch size of 8, learning rate of 1e-05 and 3 epochs.

F. EXPERIMENTAL RESULTS

This section is dedicated to the experiments. First, we conducted many experiments for choosing the best BERT pre-trained model (different versions of AraBERT, BERT-Base Multilingual-cased and uncased) to map tweets to the contextualized embeddings. Second, we compared the performance of the proposed model with state-of-the-art and baseline Arabic emotion detection methods. Then, another comparison will be made with the top performers of deep learning models, namely, LSTM, LSTM - BiLSTM, and our proposed approach (Attentional LSTM - BiLSTM). Finally, to boost and discuss more experiments, we fine-tuned two versions of BERT-Base Multilingual (Cased and Uncased) and different versions of AraBERT with and without segmentation for the emotion detection task on the SemEval2018-Ar dataset.

In the first experiment, the proposed method was used with two versions of BERT-Base Multilingual (Cased and Uncased) and with different versions of AraBERT as the language-specific pre-trained model utilized for mapping each tweet into the corresponding contextualized embeddings. As shown in Table 8, all versions of AraBERT perform slightly better than BERT-Base Multilingual. Furthermore, AraBERT-v2-large performs better than AraBERT-v1-base and AraBERT-v2-base, which can be explained by the high number of total parameter tunings (371 M for AraBERT-v2-large compared to the 100+M for the other AraBERT models' size), the number of transformer layers, the number of hidden units in each layer, the number of attention heads per hidden unit, and the size of the vocabulary, as discussed previously in Table 4. Furthermore, the results are much better without segmentation, which is related to the segmentation that does not perform well on Dialectal Arabic (DA) and noisy data of SemEval2018-Ar Twitter dataset since AraBERT was trained on Modern Standard Arabic (MSA), which is found in today's written scripts and spoken mainly in formal channels.

Table 9 presents the comparison results between the proposed approach and the state-of-the-art Arabic emotion analysis methods on the SemEval2018-Ar dataset. We notice that our deep attentional LSTM-BiLSTM model outperforms the results reported in SemEval2018-Task1: (Affect in Tweets) competition with an enhancement of 4.92% over the best performing model (i.e., EMA [43]). The majority of the reported works shown in Table 9 have participated in SemEval2018-Task1: (Affect in Tweets) competition. Being competition-based research, the members

TABLE 7. Summary of hyper-parameters.

Name	*Details*
Optimizer	RMSProp optimizer [77]
Learning Rate	0.001
Back-propagation	ReLu
Batch size	4
Dropout	0.25
L2 regularization	0.001
Hidden layer dimension	128 each

[4]https://colab.research.google.com/
[5]https://keras.io
[6]https://huggingface.co/

TABLE 8. Comparison with two versions of BERT-base multilingual (cased and uncased) and with different versions of AraBERT model.

BERT model	Accuracy (%)	F1-measure (%)
BERT-Base Multilingual- Cased	20.1	19.2
BERT-Base Multilingual- Uncased	25.3	22.3
AraBERTv01-base-without Pre-Segmentation	48.55	31.26
AraBERTv01-base-with Pre-Segmentation	39.83	20.93
AraBERTv02-base-without Pre-Segmentation	51.43	24.02
AraBERTv02-base-with Pre-Segmentation	50.62	22.87
AraBERTv02-large-without Pre-Segmentation	**53.82**	**32.04**
AraBERTv02-large-with Pre-Segmentation	53.33	27.60

TABLE 9. The proposed approach accuracy (%) on emotion analysis task compared to other related work on SemEval2018-AR dataset.

Approach	Accuracy (%)
Random Baseline [31]	17.70
MEDIAN Team [31]	25.40
SVM-Unigrams [31]	38.00
TeamUNCC [42]	44.60
Tw-StAR [44]	46.50
PARTNA [31]	48.40
EMA [43]	48.90
Alswaidan et al., 2020 [21]	51.20
Our approach (Attentional LSTM - BiLSTM)	**53.82**

develop computationally expensive models to accomplish higher results. For example, EMA [43], PARTNA [31], and Tw-StAR [44] are the top based models.

In addition, on the SemEval2018-Ar dataset, our proposed model outperforms the current state-of-the-art Alswaidan and Menai [21] model, achieving 2.62% improvement in accuracy. To the best of our knowledge, our model outperforms the best recently reported model in the literature.

TABLE 10. Arabic emotion classification using top performers of the deep learning models.

Approach	Accuracy (%)	F1-measure (%)
LSTM	41.02	21.42
LSTM - BiLSTM	48.72	22.95
Our approach (LSTM - BiLSTM with attention mechanism)	**53.82**	**32.04**

The third set of experiments is dedicated to multilabel Arabic emotion classification using the top performers of the deep learning models, namely, LSTM, LSTM-BiLSTM, and our attentional LSTM- BiLSTM. The objective of this set of experiments was to tag each tweet in the test dataset with 'neutral or no emotion' or as one, or more, of eleven given emotions (*anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust*) that best represent the mental state of the tweeter. Having a closer look at the results in Table 10, our LSTM-BiLSTM with attention mechanism achieved competitive results.

To boost and discuss more experiments, we fine-tuned the transformer BERT model for the Arabic multilabel emotion detection task on the SemEval2018-Ar dataset. We fine-tuned two versions of BERT-Base Multilingual (Cased and Uncased) and different versions of pre-trained AraBERT (AraBERTv01, AraBERTv02) with and without segmentation.

Table 11 presents the results. All versions of AraBERT perform slightly better than BERT-Base Multilingual. We observed that AraBERTv2-large fine-tuning can detect 61% of emotion labels (micro F1-measure).

To gain insight into the performance of our system, we calculated the AUC of each label. The results of this analysis are shown in Figure 6. We observe that our system based on AraBERTv2 consistently outperforms the other variants of AraBERT and BERT-Multilingual in all emotion labels. Furthermore, we notice that the system based on AraBERTv02-large gave the best performance on the "joy" label followed by the "anger", "sadness", and "fear" labels. The worst performance was obtained on the "surprise", "trust", and "anticipation" labels. The reason could be the low number of training examples (Table 6) containing these emotion labels ("surprise": 47, "trust": 120, and "anticipation": 209) and the out-of-vocabulary (OOV) issue.

As the dataset used in our experiment is relatively small, the models based on deep learning may suffer from overfitting during the training phase. Figure 7 depicts the comparison between the training and validation loss values computed at the end of each training epoch, showing that our model was

TABLE 11. Results performance of fine-tuning two versions of BERT-base multilingual (cased and uncased) and fine-tuning different versions of AraBERT model.

Model		Average	Precision	Recall	F1-measure
BERT-base-multilingual-cased		Micro	0.64	0.26	0.37
		Macro	0.84	0.18	0.20
BERT-base-multilingual-uncased		Micro	0.62	0.29	0.40
		Macro	0.84	0.20	0.24
AraBERT v01-base	Without pre-Segmentation	Micro	0.75	0.50	0.60
		Macro	0.85	0.37	0.40
	With pre-Segmentation	Micro	0.66	0.44	0.53
		Macro	0.86	0.30	0.32
AraBERT v02-base	Without pre-Segmentation	Micro	0.74	0 .50	0.59
		Macro	0.85	0.35	0.36
	With pre-Segmentation	Micro	0.73	0.44	0.55
		Macro	0.85	0.31	0.33
AraBERT v02-large	Without pre-Segmentation	**Micro**	**0.79**	**0.50**	**0.61**
		Macro	**0.90**	**0.41**	**0.43**
	With pre-Segmentation	Micro	0.78	0.49	0.60
		Macro	0.88	0.40	0.42

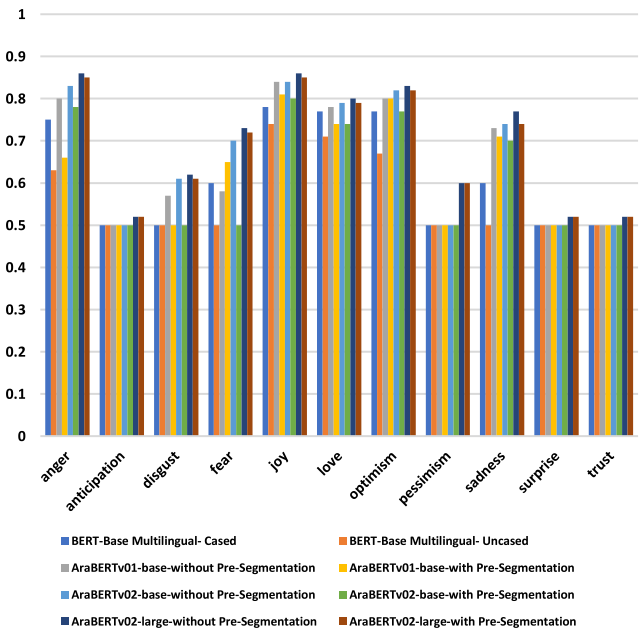

FIGURE 6. AUC performance per label in SemEval2018-Ar E-c task.

trained without overfitting. Furthermore, Figure 8 visualizes the word cloud of most common words in (a) Angry, (b) Joy, (c) Sadness, and (d) Love.

1) ERROR ANALYSIS

We provide both quantitative and qualitative analysis to showcase the strength and weakness of the proposed approach.

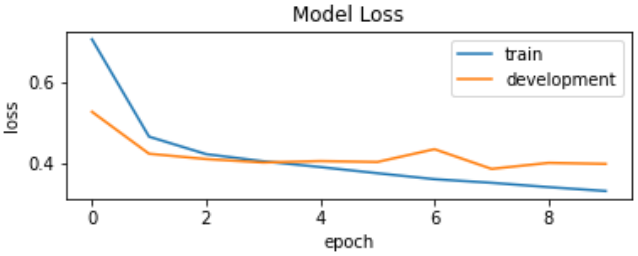

FIGURE 7. Our model loss plots for training and validation data.

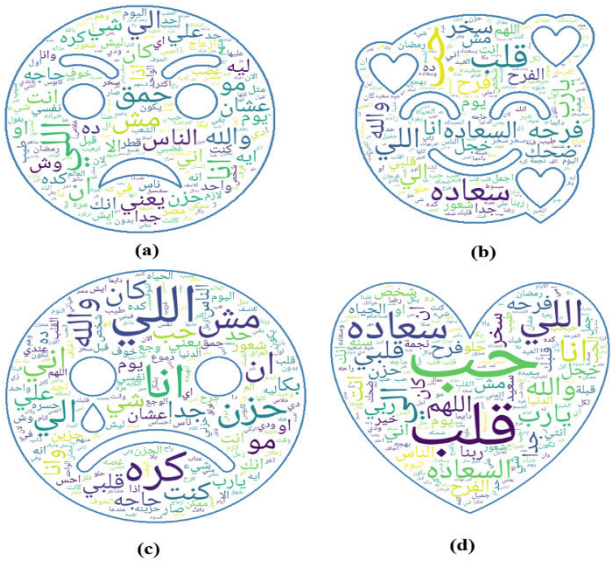

FIGURE 8. Word cloud of (a) Angry, (b) Joy, (c) Sadness, and (d) Love.

Figure 9 shows the confusion matrix of the proposed model on the development set. We notice that using the contextualized embeddings in the Arabic affect analysis task, the classifier is able to distinguish better amongst emotion classes. Hence, we can determine that the subtlety of emotion is better learned after combining the pre-trained contextualized embeddings and the proposed attentional LSTM-BiLSTM deep model. Furthermore, concerning the labels ''anger'', ''love'', ''optimism'', and ''sadness'', the confusion matrices show respectively that 142 of 505 tweets were predicted as false ''anger'' (FN), 175 of 585 tweets were predicted as false ''love'' (FN), 169 of 585 tweets were predicted as false ''optimism'' (FN), and 159 of 505 were predicted as false ''sadness'' (FN). One reason may be that there is a positive correlation between these emotions in training examples: optimism-joy-love, sadness-pessimism, and anger-disgust, as shown in Figure 5. We have an intuition to overcome this drawback by investigating data-object properties to identify constraints of conjunctions of positive and negative semantics using highly comprehensive knowledge such as SenticNet6 [78].

G. DISCUSSION

The purpose of this work is to propose an affect analysis approach tailored to Arabic tweets. The experimental results

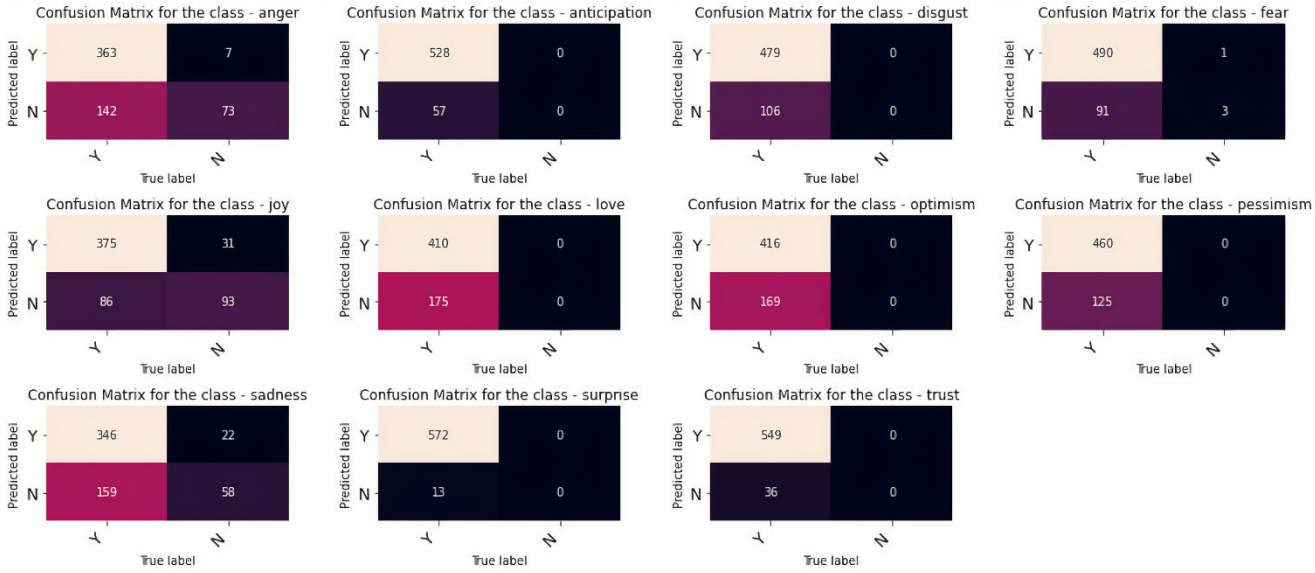

FIGURE 9. The confusion matrices of the proposed model on the development set.

show that our proposal outperforms the current state-of-the-art methods. This improvement can be explained by the following reasons: (i) we have enriched the tweets by transcribing their embedded emojis to their corresponding Arabic words, (ii) the contextualized embeddings are captured by the AraBERT pre-trained model, and (iii) the proposed attention-based LSTM-BiLSTM deep model determines the label-emotion of tweets.

All versions of AraBERT perform slightly better than BERT-Base Multilingual, indicating that language-specific pretraining has improved the performance of the proposed method. Eventually, the AraBERT-v2 model still shows the highest results compared to AraBERT-v1 because AraBERT-v2 is specifically trained on more Arabic data and better vocabulary. Additionally, AraBERT-v2, as a larger model, outperforms smaller models pre-trained on the language-specific text.

Furthermore, the attention-based LSTM-BiLSTM deep model for affect classification achieves a considerable gain compared to other sequential deep learning models, namely, LSTM, and LSTM-BiLSTM. This is because the attentional LSTM-BiLSTM model can more effectively learn the context of each word in the tweet and capture the most significant part of a target sentence. Therefore, this trick improves the performance of multilabel emotion classification.

However, some limitations have been identified. Our system does not perform well with ''sadness'', ''surprise'', and ''anticipation'' emotion labels, which can be explained by the low number of training examples related to these emotion labels. Thus, in our future work, we plan to overcome this drawback. One possible solution is that we are most likely to work more on ways to use transfer learning.

VI. CONCLUSION AND FUTURE WORK

In this work, we have addressed the affect analysis problem for Arabic tweets. We have proposed an approach that combines AraBERT to generate the contextualized embeddings of Arabic tweets and an attentional LSTM-BiLSTM as a multilabel emotion classification model. Experiments are conducted on the reference dataset SemEval-2018 Task1. The comprehensive results show that our proposed approach outperforms eight current state-of-the-art methods and baseline methods. It achieves significant accuracy (53.82%) compared to 1st place (48.9%) in the SemEval2018-Task1: (Affect in Tweets) competition. Additionally, it outperforms the best recently reported model in the literature [21] with an enhancement of 2.62% in accuracy on the SemEval2018-Ar dataset. We noticed that investigating deep contextualized language models can significantly improve the performance of Arabic affect analysis.

Furthermore, the current work can provide many benefits for governments, health authorities, and decision-makers to monitor people's emotions on top of social media content. Additionally, our current work is designed to improve business strategies according to the emotions of customers and recognize potential criminals when analysing the emotions of people after an attack or crime.

Another point is worth mentioning in the long term, the pandemic caused by COVID-19 led to the spread of excessive pseudoscientific information and fake news that confused public health status. For future work, we plan to build a web-based emotion recognizer able to crawl tweets, filter fake news and misleading information and then detect the emotion label in real-time. The system can be helpful for recognizing and analysing people's emotions during any future epidemic.

MODE OF AVAILABILITY

The python source code of the proposed system is available at https://colab.research.google.com/drive/1kfHnNVZ0zs4zEahzZtBAKnlH8ybUInzz?usp=sharing

REFERENCES

[1] K. Sailunaz, M. Dhaliwal, J. Rokne, and R. Alhajj, "Emotion detection from text and speech: A survey," *Soc. Netw. Anal. Mining*, vol. 8, no. 1, p. 28, 2018.

[2] F. Kalloubi, N. El Habib, and O. El Beqqali, "Graph based tweet entity linking using DBpedia," in *Proc. IEEE/ACS 11th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Nov. 2014, pp. 501–506.

[3] S. Bashir, S. Bano, S. Shueb, S. Gul, A. A. Mir, R. Ashraf, and N. Noor, "Twitter chirps for Syrian people: Sentiment analysis of tweets related to Syria chemical attack," *Int. J. Disaster Risk Reduction*, vol. 62, Aug. 2021, Art. no. 102397, doi: 10.1016/j.ijdrr.2021.102397.

[4] K. Garcia and L. Berton, "Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA," *Appl. Soft Comput.*, vol. 101, Mar. 2021, Art. no. 107057, doi: 10.1016/j.asoc.2020.107057.

[5] D. Xu, Z. Tian, R. Lai, X. Kong, Z. Tan, and W. Shi, "Deep learning based emotion analysis of microblog texts," *Inf. Fusion*, vol. 64, pp. 1–11, Dec. 2020, doi: 10.1016/j.inffus.2020.06.002.

[6] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowl.-Based Syst.*, vol. 226, Aug. 2021, Art. no. 107134.

[7] F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah, "Text-based emotion detection: Advances, challenges, and opportunities," *Eng. Rep.*, vol. 2, no. 7, Jul. 2020, Art. no. e12189.

[8] H. S. Ibrahim, S. M. Abdou, and M. Gheith, "MIKA: A tagged corpus for modern standard Arabic and colloquial sentiment analysis," in *Proc. IEEE 2nd Int. Conf. Recent Trends Inf. Syst. (ReTIS)*, Jul. 2015, pp. 353–358.

[9] M. Korayem, D. Crandall, and M. Abdul-Mageed, "Subjectivity and sentiment analysis of Arabic: A survey," in *Proc. Int. Conf. Adv. Mach. Learn. Technol. Appl.*, Dec. 2012, pp. 128–139.

[10] M. Al-Ayyoub, A. A. Khamaiseh, Y. Jararweh, and M. N. Al-Kabi, "A comprehensive survey of Arabic sentiment analysis," *Inf. Process. Manag.*, vol. 56, no. 2, pp. 320–342, Mar. 2019, doi: 10.1016/j.ipm.2018.07.006.

[11] G. Badaro, R. Baly, H. Hajj, W. El-Hajj, K. B. Shaban, N. Habash, A. Al-Sallab, and A. Hamdi, "A survey of opinion mining in Arabic: A comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 18, no. 3, pp. 1–52, Jul. 2019, doi: 10.1145/3295662.

[12] A. Alsayat and N. Elmitwally, "A comprehensive study for Arabic sentiment analysis (challenges and applications)," *Egyptian Informat. J.*, vol. 21, no. 1, pp. 7–12, Mar. 2020, doi: 10.1016/j.eij.2019.06.001.

[13] O. Oueslati, E. Cambria, M. B. HajHmida, and H. Ounelli, "A review of sentiment analysis research in Arabic language," *Future Gener. Comput. Syst.*, vol. 112, pp. 408–430, Nov. 2020, doi: 10.1016/j.future.2020.05.034.

[14] A. B. Nassif, A. Elnagar, I. Shahin, and S. Henno, "Deep learning for Arabic subjective sentiment analysis: Challenges and research opportunities," *Appl. Soft Comput.*, vol. 98, Jan. 2021, Art. no. 106836, doi: 10.1016/j.asoc.2020.106836.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: http://arxiv.org/abs/1301.3781

[16] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[17] J. Camacho-Collados and M. T. Pilevar, "From word to sense embeddings: A survey on vector representations of meaning," *J. Artif. Intell. Res.*, vol. 63, pp. 743–788, Dec. 2018.

[18] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," 2020, *arXiv:2003.00104*. [Online]. Available: http://arxiv.org/abs/2003.00104

[19] K. S. Kalaivani, S. Uma, and C. S. Kanimozhiselvi, "A review on feature extraction techniques for sentiment classification," in *Proc. 4th Int. Conf. Comput. Methodologies Commun. (ICCMC)*, Mar. 2020, pp. 679–683, doi: 10.1109/ICCMC48092.2020.ICCMC-000126.

[20] H. Liang, X. Sun, Y. Sun, and Y. Gao, "Text feature extraction based on deep learning: A review," *EURASIP J. Wireless Commun. Netw.*, vol. 2017, no. 1, pp. 1–12, Dec. 2017.

[21] N. Alswaidan and M. E. B. Menai, "Hybrid feature model for emotion recognition in Arabic text," *IEEE Access*, vol. 8, pp. 37843–37854, 2020.

[22] U. Naseem, I. Razzak, K. Musial, and M. Imran, "Transformer based deep intelligent contextual embedding for Twitter sentiment analysis," *Future Gener. Comput. Syst.*, vol. 113, pp. 58–69, Dec. 2020.

[23] M. E. Basiri, S. Nemati, M. Abdar, S. Asadi, and U. R. Acharrya, "A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets," *Knowl.-Based Syst.*, vol. 228, Sep. 2021, Art. no. 107242, doi: 10.1016/j.knosys.2021.107242.

[24] I. Abu Farha and W. Magdy, "A comparative study of effective approaches for Arabic sentiment analysis," *Inf. Process. Manage.*, vol. 58, no. 2, Mar. 2021, Art. no. 102438, doi: 10.1016/j.ipm.2020.102438.

[25] S. Al-Dabet, S. Tedmori, and M. AL-Smadi, "Enhancing Arabic aspect-based sentiment analysis using deep learning models," *Comput. Speech Lang.*, vol. 69, Sep. 2021, Art. no. 101224.

[26] A. R. Pathak, M. Pandey, and S. Rautaray, "Topic-level sentiment analysis of social media data using deep learning," *Appl. Soft Comput.*, vol. 108, Sep. 2021, Art. no. 107440, doi: 10.1016/j.asoc.2021.107440.

[27] O. AlZoubi, S. K. Tawalbeh, and M. Al-Smadi, "Affect detection from Arabic tweets using ensemble and deep learning techniques," *J. King Saud Univ.-Comput. Inf. Sci.*, 2020, doi: 10.1016/j.jksuci.2020.09.013.

[28] A. M. Nerabie, M. AlKhatib, S. S. Mathew, M. E. Barachi, and F. Oroumchian, "The impact of Arabic part of speech tagging on sentiment analysis: A new corpus and deep learning approach," *Procedia Comput. Sci.*, vol. 184, pp. 148–155, Jan. 2021, doi: 10.1016/j.procs.2021.03.026.

[29] A. E. de Oliveira Carosia, G. P. Coelho, and A. E. A. da Silva, "Investment strategies applied to the Brazilian stock market: A methodology based on sentiment analysis with deep learning," *Expert Syst. Appl.*, vol. 184, Dec. 2021, Art. no. 115470, doi: 10.1016/j.eswa.2021.115470.

[30] H. Elfaik and E. H. Nfaoui, "Deep bidirectional LSTM network learning-based sentiment analysis for Arabic text," *J. Intell. Syst.*, vol. 30, no. 1, pp. 395–412, Dec. 2020, doi: 10.1515/jisys-2020-0021.

[31] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 task 1: Affect in tweets," in *Proc. 12th Int. Workshop Semantic Eval.*, Jun. 2018, pp. 1–17.

[32] A. F. El Gohary, T. I. Sultan, M. A. Hana, and M. M. El Dosoky, "A computational approach for analyzing and detecting emotions in Arabic text," *Int. J. Eng. Res. Appl.*, vol. 3, pp. 100–107, May 2013.

[33] O. Rabie and C. Sturm, "Feel the heat: Emotion detection in Arabic social media content," in *Proc. Int. Conf. Data Mining, Internet Comput., Big Data (BigData)*, Nov. 2014, pp. 37–49.

[34] M. Al-A'abed and M. Al-Ayyoub, "A lexicon-based approach for emotion analysis of Arabic social media content," in *Proc. Int. Comput. Sci. Inform. Conf. (ICSIC)*, 2016, pp. 343–351.

[35] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets," 2013, *arXiv:1308.6242*. [Online]. Available: https://arxiv.org/abs/1308.6242

[36] W. A. Hussien, Y. M. Tashtoush, M. Al-Ayyoub, and M. N. Al-Kabi, "Are emoticons good enough to train emotion classifiers of Arabic tweets?" in *Proc. 7th Int. Conf. Comput. Sci. Inf. Technol. (CSIT)*, Jul. 2016, pp. 1–6.

[37] A. Al-Khatib and S. R. El-Beltagy, "Emotional tone detection in Arabic tweets," in *Proc. Int. Conf. Comput. Linguistics Intell. Text Process.*, Apr. 2017, pp. 105–114.

[38] A. M. Abd Al-Aziz, M. Gheith, and A. S. Eldin, "Lexicon based and multi-criteria decision making (MCDM) approach for detecting emotions from Arabic microblog text," in *Proc. 1st Int. Conf. Arabic Comput. Linguistics (ACLing)*, Apr. 2015, pp. 100–105.

[39] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. Al-Ayyoub, "Arabic sentiment analysis: Lexicon-based and corpus-based," in *Proc. IEEE Jordan Conf. Appl. Electr. Eng. Comput. Technol. (AEECT)*, Dec. 2013, pp. 1–6.

[40] O. Badarneh, M. Al-Ayyoub, N. Alhindawi, L. A. Tawalbeh, and Y. Jararweh, "Fine-grained emotion analysis of Arabic tweets: A multi-target multi-label approach," in *Proc. IEEE 12th Int. Conf. Semantic Comput. (ICSC)*, Jan. 2018, pp. 340–345.

[41] A. M. Sayed, S. AbdelRahman, R. Bahgat, and A. Fahmy, "Time emotional analysis of Arabic tweets at multiple levels," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 10, pp. 336–342, 2016.

[42] M. Abdullah and S. Shaikh, "TeamUNCC at SemEval-2018 task 1: Emotion detection in English and Arabic tweets using deep learning," in *Proc. 12th Int. Workshop Semantic Eval.*, Jun. 2018, pp. 350–357.

[43] G. Badaro, O. El Jundi, A. Khaddaj, A. Maarouf, R. Kain, H. Hajj, and W. El-Hajj, "EMA at SemEval-2018 task 1: Emotion mining for Arabic," in *Proc. 12th Int. Workshop Semantic Eval.*, 2018, pp. 236–244.
[44] H. Mulki, C. B. Ali, H. Haddad, and I. Babaoglu, "Tw-StAR at SemEval-2018 task 1: Preprocessing impact on multi-label emotion classification," in *Proc. 12th Int. Workshop Semantic Eval.*, 2018, pp. 167–171.
[45] M. Abdullah, M. Hadzikadicy, and S. Shaikhz, "SEDAT: Sentiment and emotion detection in Arabic text using CNN-LSTM deep learning," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2018, pp. 835–840.
[46] M. Jabreel and A. Moreno, "EiTAKA at SemEval-2018 task 1: An ensemble of N-channels ConvNet and XGboost regressors for emotion analysis of tweets," 2018, *arXiv:1802.09233*. [Online]. Available: http://arxiv.org/abs/1802.09233
[47] S. K. Tawalbehe, O. AlZoubi, and M. AL-Smadi, "Recent advances of affect detection from Arabic text," in *Proc. 10th Int. Conf. Inf. Commun. Syst. (ICICS)*, Jun. 2019, pp. 128–133.
[48] S. Lalitha, S. Patnaik, T. H. Arvind, V. Madhusudhan, and S. Tripathi, "Emotion recognition through speech signal for human-computer interaction," in *Proc. 5th Int. Symp. Electron. Syst. Design*, Dec. 2014, pp. 217–218.
[49] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," in *Proc. 8th Eur. Conf. Speech Commun. Technol.*, 2003, pp. 125–128.
[50] R. A. Calvo and S. M. Kim, "Emotions in text: Dimensional and categorical models," *Comput. Intell.*, vol. 29, no. 3, pp. 527–543, Aug. 2013.
[51] I. Perikos and I. Hatzilygeroudis, "Recognizing emotions in text using ensemble of classifiers," *Eng. Appl. Artif. Intell.*, vol. 51, pp. 191–201, May 2016.
[52] A. F. M. N. H. Nahin, J. M. Alam, H. Mahmud, and K. Hasan, "Identifying emotion by keystroke dynamics and text pattern analysis," *Behaviour Inf. Technol.*, vol. 33, no. 9, pp. 987–996, Sep. 2014.
[53] S. M. Mohammad and S. Kiritchenko, "Using hashtags to capture fine emotion categories from tweets," *Comput. Intell.*, vol. 31, no. 2, pp. 301–326 May 2015.
[54] I. Cohen, A. Garg, and T. S. Huang, "Emotion recognition from facial expressions using multilevel HMM," in *Proc. Neural Inf. Process. Syst.*, vol. 2, 2000, pp. 1–7.
[55] M. Rosenblum, Y. Yacoob, and L. S. Davis, "Human expression recognition from motion using a radial basis function network architecture," *IEEE Trans. Neural Netw.*, vol. 7, no. 5, pp. 1121–1138, Sep. 1996.
[56] M. D. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE Trans. Affective Comput.*, vol. 5, no. 2, pp. 101–111, Apr./Jun. 2014.
[57] C. Broad, "Emotion and sentiment," *J. Aesthetics Art Criticism*, vol. 13, no. 2, pp. 203–214, 1954.
[58] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Commun.*, vol. 40, nos. 1–2, pp. 5–32, Apr. 2003.
[59] P. Ekman, "An argument for basic emotions," *Cognit. Emotion*, vol. 6, nos. 3–4, pp. 169–200, 1992.
[60] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, p. 1161, 1980.
[61] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of Emotion*, R. Plutchik and H. Kellerman, Eds. New York, NY, USA: Academic, 1980, ch. 1, pp. 3–33.
[62] S. M. Mohammad, "Sentiment analysis: Detecting valence, emotions, and other affectual states from text," in *Emotion Measurement*. Amsterdam, The Netherlands: Elsevier, 2016, pp. 201–237.
[63] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
[64] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," 2017, *arXiv:1712.09405*. [Online]. Available: http://arxiv.org/abs/1712.09405
[65] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805
[66] P. Shaver, J. Schwartz, D. Kirson, and C. O'connor, "Emotion knowledge: Further exploration of a prototype approach," *J. Pers. Soc. Psychol.*, vol. 52, no. 6, p. 1061, 1987.
[67] K. Oatley and P. N. Johnson-laird, "Towards a cognitive theory of emotions," *Cognition Emotion*, vol. 1, no. 1, pp. 29–50, Mar. 1987.
[68] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions Cambridge*. Cambridge, U.K.: Cambridge Univ. Press, 1988.
[69] H. Lövheim, "A new three-dimensional model for emotions and monoamine neurotransmitters," *Med. Hypotheses*, vol. 78, no. 2, pp. 341–348, Feb. 2012.
[70] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 19–27.
[71] M. Cliche, "BB_twtr at SemEval-2017 task 4: Twitter sentiment analysis with CNNs and LSTMs," 2017, *arXiv:1704.06125*. [Online]. Available: http://arxiv.org/abs/1704.06125
[72] H. Ghulam, F. Zeng, W. Li, and Y. Xiao, "Deep learning-based sentiment analysis for Roman Urdu text," *Procedia Comput. Sci.*, vol. 147, pp. 131–135, Jan. 2019.
[73] X. Zhang, F. Chen, and R. Huang, "A combination of RNN and CNN for attention-based relation classification," *Procedia Comput. Sci.*, vol. 131, pp. 911–917, Jan. 2018, doi: 10.1016/j.procs.2018.04.221.
[74] S. M. Liu and J.-H. Chen, "A multi-label classification based approach for sentiment classification," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1083–1093, Feb. 2015, doi: 10.1016/j.eswa.2014.08.036.
[75] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning," *Pattern Recognit.*, vol. 45, no. 9, pp. 3084–3104, Jan. 2012, doi: 10.1016/j.patcog.2012.03.004.
[76] H. W. Ian and F. Eibe, *Data Mining: Practical Machine Learning Tools and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2005.
[77] T. Tieleman and G. Hinton, "Lecture 6.5-RMSPROP: Divide the gradient by a running average of its recent magnitude," *COURSERA, Neural Netw. Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.
[78] E. Cambria, Y. Li, F. Z. Xing, S. Poria, and K. Kwok, "SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2020, pp. 105–114.

HANANE ELFAIK was born in Morocco, in 1992. She received the master's degree in computer science from the Faculty of Science Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco, where she is currently pursuing the Ph.D. degree. Her research interests include sentiment analysis, text mining, deep learning, and natural language processing.

EL HABIB NFAOUI (Member, IEEE) received the Ph.D. degree in computer science from Sidi Mohamed Ben Abdellah University, Fez, Morocco, and the University of Lyon, France, under a cotutelle agreement (doctorate in joint supervision), in 2008, and the HU Diploma degree (accreditation to supervise research) in computer science from Sidi Mohamed Ben Abdellah University, in 2013. He is currently a Professor of computer science with Sidi Mohamed Ben Abdellah University. He has published in international reputed journals, books, and conferences, and has edited seven conference proceedings and special issue books. His current research interests include information retrieval, language representation learning, machine learning and deep learning, web mining and text mining, semantic web, web services, social networks, and multi-agent systems. He is a Co-Founder and an Executive Member of the International Neural Network Society Morocco Regional Chapter. He is also a Co-Founder and the Chair of the IEEE Morocco Section Computational Intelligence Society Chapter. He has co-founded the International Conference on Intelligent Computing in Data Sciences (ICSD2017) and the International Conference on Intelligent Systems and Computer Vision (ISCV2015). He has served as a reviewer for scientific journals and on the program committee for several conferences.