

An Ensemble Deep Learning Approach for Emotion Detection in Arabic Tweets

Alaa Mansy, Sherine Rady, Tarek Gharib

Department of Information Systems, Faculty of Computers and Information, Ain Shams University, Cairo, Egypt

Abstract—Now-a-days people use social media websites for different activities such as business, entertainment, following the news, expressing their thoughts, feelings, and much more. This initiated a great interest in analyzing and mining such user-generated content. In this paper, the problem of emotion detection (ED) in Arabic text is investigated by proposing an ensemble deep learning approach to analyze user-generated text from Twitter, in terms of the emotional insights that reflect different feelings. The proposed model is based on three state-of-the-art deep learning models. Two models are special types of Recurrent Neural Networks (Bi-LSTM and Bi-GRU), and the third model is a pre-trained language model (PLM) based on BERT and it is called MARBERT transformer. The experiments were evaluated using the SemEval-2018-Task1-Ar-Ec dataset that was published in a multilabel classification task: Emotion Classification (EC) inside the SemEval-2018 competition. MARBERT PLM is compared to one of the most famous PLM for dealing with the Arabic language (AraBERT). Experiments proved that MARBERT achieved better results with an improvement of 4%, 2.7%, 4.2%, and 3.5% regarding Jaccard accuracy, recall, F1 macro, and F1 micro scores respectively. Moreover, the proposed ensemble model showed outperformance over the individual models (Bi-LSTM, Bi-GRU, and MARBERT). It also outperforms the most recent related work with an improvement ranging from 0.2% to 4.2% in accuracy, and from 5.3% to 23.3% in macro F1 score.

Keywords—Deep learning; emotion detection; transformers; RNNs; Bi-LSTM; Bi-GRU

I. INTRODUCTION

Twitter is a famous 24/7 active social media platform with many signed users sharing their activities, thoughts, and feelings at any time. People post tweets, stream live videos, chat with each other, companies create and manage a lot of marketing campaigns to promote their products, and even much more services are provided.

These days no one can give up using online social networks because it makes them feel connected all the time. Also, they can express their feelings and emotions whether they are happy, sad, surprised, anticipated, or any other feelings during their online activities.

A lot of expressions and words in our daily written text over the web may reflect our feelings. Not only that but also it may affect other people significantly because we believe that every simple word reflects an impact. For example, posting a tweet like that:

“I got COVID-19 twice even though I have been vaccinated the vaccine is useless”,

Such simple words can kill a lot of people affected by that virus. Elderly people who have chronic diseases will realize that death is their next step because it makes them feel frustrated. By analyzing that content, everything that may affect a lot of people can be controlled. For Example, social networks can utilize a model for emotion detection in their platforms as an option to prevent such disappointing statuses from being appeared in their customers' timelines. In this way, they can control and restrict anxiety, frustration, and much more.

Emotion Detection or ED is one of the hottest research topics in the field of Natural Language Processing (NLP). ED is considered different from Sentiment Analysis (SA), where SA task is to recognize polarities from text such as positive, negative, or neutral. On the other hand, ED aims to get emotional insights from what has been typed [1].

When reading a tweet, it may reflect one of the following feeling polarities (positive, negative, or neutral). This study is not focusing on the detection of these polarities, but it goes deeper to detect different emotions like (joy, anger, surprise, etc.).

Scientists have summarized ED activities in a set of approaches that determine how exactly emotions are represented. The most famous emotional model is the Discrete Emotion Model (DEM) like Ekman's model which contains six basic emotions which are anger, fear, disgust, happiness, sadness, and surprise. The other models are Dimensional Emotion Model (DiEM) like Plutchik's Emotion Model and Russell's Circumplex Model [2][3].

Suppose a text presented in a user-generated tweet like this: “غضب دفين يفقد الاشياء الوانها لتصبح رمادية وتنعدم لذة الحياة”, humans can simply understand the context of this sentence by understanding each word based on the understanding of previous and next words. Also, they can understand the implied emotion of the user who posted the tweet (tweeter) which is sad or angry.

Words in a sentence are linked with each other's in a certain sequence to form a meaning, understanding that meaning is called “Contextual Understanding”. Traditional machine learning techniques cannot understand the context very well. Deep learning (DL) sequence models can be utilized to make machines simulate human understanding.

Sequence models such as Recurrent Neural Networks (RNNs) can understand the context by memorizing words and getting the relationships between them. But they have shown some shortages known by the problems of vanishing and

exploding gradients. Accordingly, new generations of RNNs have been developed to overcome that shortage. For example, LSTM and GRU models can deal with long-term dependencies and tackle the problems mentioned above.

Although there is a lack of Arabic resources and research studies on Arabic contextual understanding, different Arabic language models were developed to support this point. The most famous one is called AraBERT [4] which is an Arabic pre-trained language model (PLM) based on Bidirectional Encoder Representations from Transformers (BERT) [5]. AraBERT was pretrained using more than 20GB of Arabic text from different sources like Arabic Wikipedia, Arabic news websites, and others. It is based on Modern Standard Arabic (MSA), and it gets better results when fine-tuned using MSA datasets. PLMs are considered a part of Transfer Learning (TL) that support the research in this area and tackle the problem of limited resources. Other models were pretrained based on both MSA and Arabic Dialects (AD) like MARBERT PLM that is utilized in this study.

State-of-the-art neural networks (Bi-LSTM, Bi-GRU, and MARBERT) have been ensembled to deal with a multilabel classification task for emotion detection in user-generated Arabic tweets that were collected and shared during SemEval-2018 task-1: Affect in Tweets.

In Section II, related work is discussed. Section III discusses the proposed ensemble model. In Section IV, empirical results and discussion are investigated. Finally in Section V, the conclusion and future work.

II. RELATED WORK

A lot of research studies have been conducted to get emotional insights and understand the context of English text. Unfortunately, there exist few studies related to the Arabic language because of different challenges related to the complexity of this language, the lack of existing Arabic resources, and different available Arabic dialects. Attention to the analysis of the Arabic language has increased in the last decade due to the need for digital transformation in Arab communities. The Arabic language has a lot of different dialects that are spoken by around 422 million speakers all over the world which is considered a big challenge in the analysis. In the following subsections, ED generic and closely related studies are discussed.

A. Survey Studies

Alswaidan et al. [6] surveyed the state-of-the-art approaches related to emotion detection ED in the textual content for English and some other languages. They mentioned the available resources (corpora and lexicons) for working with ED tasks and addressed some challenges like (I) The challenge of detecting implicit emotions which are hidden in the text. (II) The problems related to size and quality in the available datasets. (III) Limited resources in some languages like Arabic.

Another survey study by Acheampong et al. [7] investigated the ED problem in text content by mentioning all available emotion-related datasets like (ISEAR, SemEval, EMOBANK, EmoInt, Cecilia Ovesdotter Alm's Affect data,

Daily Dialog, AMAN'S Emotion, Grounded Emotion data, Emotion-Stimulus data, Crowdsourcing, MELD, Emotion and Smile dataset). Also, they have mentioned the different approaches used to analyze and detect the emotional insights from that data (the rule construction approach, ML approach, and the hybrid approach). And they have made a comparison between different related works in terms of (used approaches, datasets, and limitations).

Similarly, a Systematic Literature Review (SLR) was introduced by William et al. [8] and listed the closely related studies used for text-based depression detection. Also, they aimed to identify and analyze different text-based approaches for the early detection of depression in social media posts. Their results showed that using BiLSTM along with the attention model performs well on depression-related textual data. They also made an experiment by using a BERT-based model and achieved better results compared to the studies mentioned in the SLR. Their experiments used a BERT-based model for the classification task. They also suggested a new method to deal with long sequences by summarizing the text before feeding it into the model. The model depends on a dataset crawled from Reddit.

B. Utilizing Traditional ML and DL Models

Mohammad et al. [9] shared a task called "Affect in Tweets" in the SemEval-2018 competition, which includes a list of subtasks for detecting the emotional states of the tweeters from their text-based tweets. They streamed and annotated some Arabic tweets to form twitter-based labeled datasets represented in three different languages English, Arabic, and Spanish. About 200 team members participated in this competition. Different ML and DL algorithms like (Bi-LSTM, CNN, Gradient Boosting, Linear Regression, Logistic Regression, LSTM, Random Forest, RNN, and SVM) were used. Badaro et al. [10] improved the performance of the emotion classification task by utilizing a pre-trained word embedding model (Aravec) and achieved the best evaluation metrics for (Arabic EC subtask) by using SVC L1 classifier that achieved 48.9%, 61.8%, 46.1% for accuracy, micro f1, and macro F1 scores respectively.

Baali et al.[11] presented a study for classifying emotions in tweets written in the Arabic language. They have used Convolutional Neural Networks (CNN) trained on top of trained word vectors. They compared the results of their approach with three ML algorithms (SVM, NB, and MLP). Their proposed approach was evaluated on the Arabic dataset provided by Sem- Eval for the emotion intensity ordinal classification task (EI-oc). Their results were 99.90% as training accuracy, and 99.82% as validation accuracy.

Khalil et al. [12] proposed a Bi-LSTM deep learning model for the task of emotion classification (EC) in Arabic tweets that were shared SemEval-2018 competition. They have merged the dataset files into only one file to use in the cross-validation process. Aravec with CBOW for the word embedding phase has been used. Their results have shown [Jaccard Accuracy 0.498, Micro Precision 0.695, Micro Recall 0.551, and Micro F1 score 0.615].

C. Utilizing Pre-Trained Language Models (PLMs)

One of the challenges of ED in Arabic text is the limited resources of the Arabic language. As a result, Transfer Learning have been emerged to help pre-train of an NLP model on one large dataset and then quickly fine-tune the model to adapt to other NLP tasks. Also, the nature of that dataset may affect the fine-tuning process i.e., if the model was pre-trained on a dataset containing emotional-related content it will give the best results in ED tasks compared to the model that was pre-trained using other natures of data. Also, some of the existing PLMs were pre-trained using Arabic MSA like AraBERT introduced by Antoun et al. [4] which gets lower results when compared to other PLMs like Abdul-Mageed et al. [13] who introduced MARBERT PLM that was pretrained using both Arabic MSA and different Arabic dialects. Another research study by Abdelali et al. [14] trained five different Arabic BERT models of QARiB using the original implementation of the BERT model implemented by google for both Arabic MSA and Arabic Dialects. Also, they have compared their results with three existing PLMs (mBERT, ARABERTv0.1&v1, ArabicBERT). And the evaluation was conducted using 5 different datasets represented in the following tasks (1) Named Entity Recognition (2) Emotion detection [SemEval2018-Ar-Ec] (3) QADI Arabic Dialects Identification (4) Offensive language detection (5) Sentiment Analysis. Macro-averaged F1 score was used as an evaluation metric, and the results related to the (EC task) using the dataset SemEval2018-Ar-Ec showed that the QARiB25 mix achieved the best macro-averaged F1 score equal to 46.8 %.

Researchers continued to investigate the development of Arabic language models thought conducting a lot of experiments like Al-Twairesh [15] who conducted ten experiments using different models starting from traditional TF-IDF to the recent state-of-the-art BERT models (TF-IDF, AraVecCBOW100, AraVecSG100, AraVecCBOW300, AraVecSG300, AraBertv01, AraBertv1, ArabicBertBase, ArabicBertLarge, Multi-Dialect Bert) on SemEval-2018 dataset. And the results showed that the Arabic BERT-Large model achieved the best results compared to other models.

Others utilized the contextualized embeddings of the PLMs to support other DL models like Elfaik et al. [2] who investigated the problem of Arabic Emotion detection (multilabel emotion classification) in tweets by combining the generated contextualized embeddings using AraBERT and an attention-based LSTM-BiLSTM deep model. The attention mechanism is applied to the output of LSTM-BiLSTM to guarantee different words. Their proposed approach was evaluated using the dataset of SemEval-2018-Task1-Ar-Ec (Affect in Tweets). Their results show that the proposed approach achieves accuracy (53.82%).

Samy et al. [16] researchers utilized some social intelligence and proposed a context-aware gated recurrent unit (C-GRU) to solve the problem of multi-label classification in Arabic-related tweets represented in the SemEval-2018-Task1-Affect in tweets (EC subtask). They have related each tweet with a specific topic, and they depend on what is called social influence where people in the same network can share topics and in the same topic, they can find similar emotions. They have used SemEval-2017 for the topic classification task and

SemEval-2018-Ec-Ar for the emotion detection task. They have used Jaccard-similarity for accuracy, F1 macro average, and F1 micro average which achieved results of 0.532, 0.648, and 0.495 respectively.

D. Utilizing Ensemble Techniques

AlZoubi et al. [1] implemented an ensemble approach that contains [bidirectional GRU_CNN (BiGRU_CNN), conventional neural networks (CNN), and XGBoost regressor (XGB)] to be used in solving the emotion intensity (EI-reg) subtask of the SemEval-2018 Task1 (Affect in Tweets). Their proposed ensemble approach was evaluated using the dataset of the SemEval-2018 Task1 EI-reg. Results show that their model achieved a Pearson of (69.2%).

Alswaidan et al. [17] proposed three different models, a human-engineered feature-based (HEF) model, a deep feature-based (DF) model, and a hybrid of both models (HEF+DF) for the emotion detection task in Arabic text. And they measured the performance of the proposed models using three different datasets (SemEval2018-Ar-Ec, IAEDS, and AETD). Regarding the SemEval2018-Ar-Ec dataset, the hybrid model achieved the best results of 0.512, 0.631, and 0.502 for Jaccard accuracy, F^{micro} , and F^{macro} scores respectively.

Talafha et al. [18] investigated the Arabic dialect identification problem and trained Arabic-BERT [19] using 10M unlabeled tweets shared in Nuanced Arabic Dialect Identification Task 1 (NADI) and the result was a new pre-trained language model called Multi-dialect-Arabic-BERT. Also, they utilized an ensemble technique (element-wise average) to get the highest value of the predicted probabilities per class for each of the four models. Their results are 44.07 for accuracy and 29.03 for the F1 score.

Closely related studies have been analyzed and concluded in the chart area shown in Fig. 1 that presents the progress till now regarding the EC task using the SemEval-2018-Ar-Ec dataset. As shown in the figure, utilizing PLMs has shown some progress in accuracy compared to other models. Similar studies that use PLMs for the EC task of SemEval-2018 didn't use the most suitable PLMs because most of them have used models that were pre-trained using non-emotional related content. To the best of our knowledge, no one has fine-tuned the MARBERT model using the SemEval-2018-Task1-Ec-Ar dataset. Also, we have used the ensemble model to combine different contextual understanding experiences that can help in getting better results.

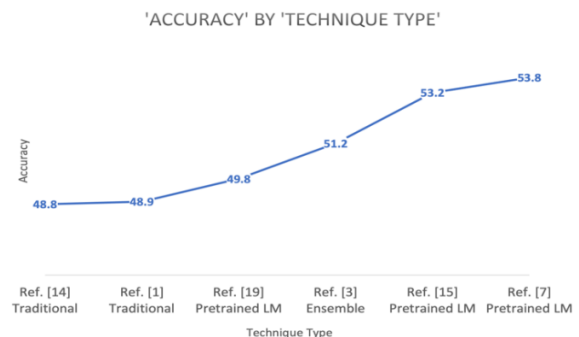


Fig. 1. Progress of Emotion Classification Task against SemEval-Ec Dataset.

III. PROPOSED MODEL

The proposed ensemble model for multi-label emotion classification EC in the Arabic language is shown in Fig. 2 which consists of six layers (a) Preprocessing layer (b) Word embedding layer (c) Processing Layer (d) Testing Layer (e) Ensemble layer (f) Classification layer. The details of these layers are explained in the following subsections.

A. Preprocessing Phase

Data preprocessing is considered one of the most important phases in machine learning applications to avoid misleading

results and get better insights. In this section, the preprocessing steps will be discussed in detail with an example from our dataset.

As shown in Table I, a user-generated tweet from the SemEval2018-Ar-Ec dataset has been preprocessed using the most common preprocessing techniques like removing English characters, numbers, stop words, repeating chars, punctuation marks, and Arabic diacritics. Also, text normalization and emojis replacement steps have been added.

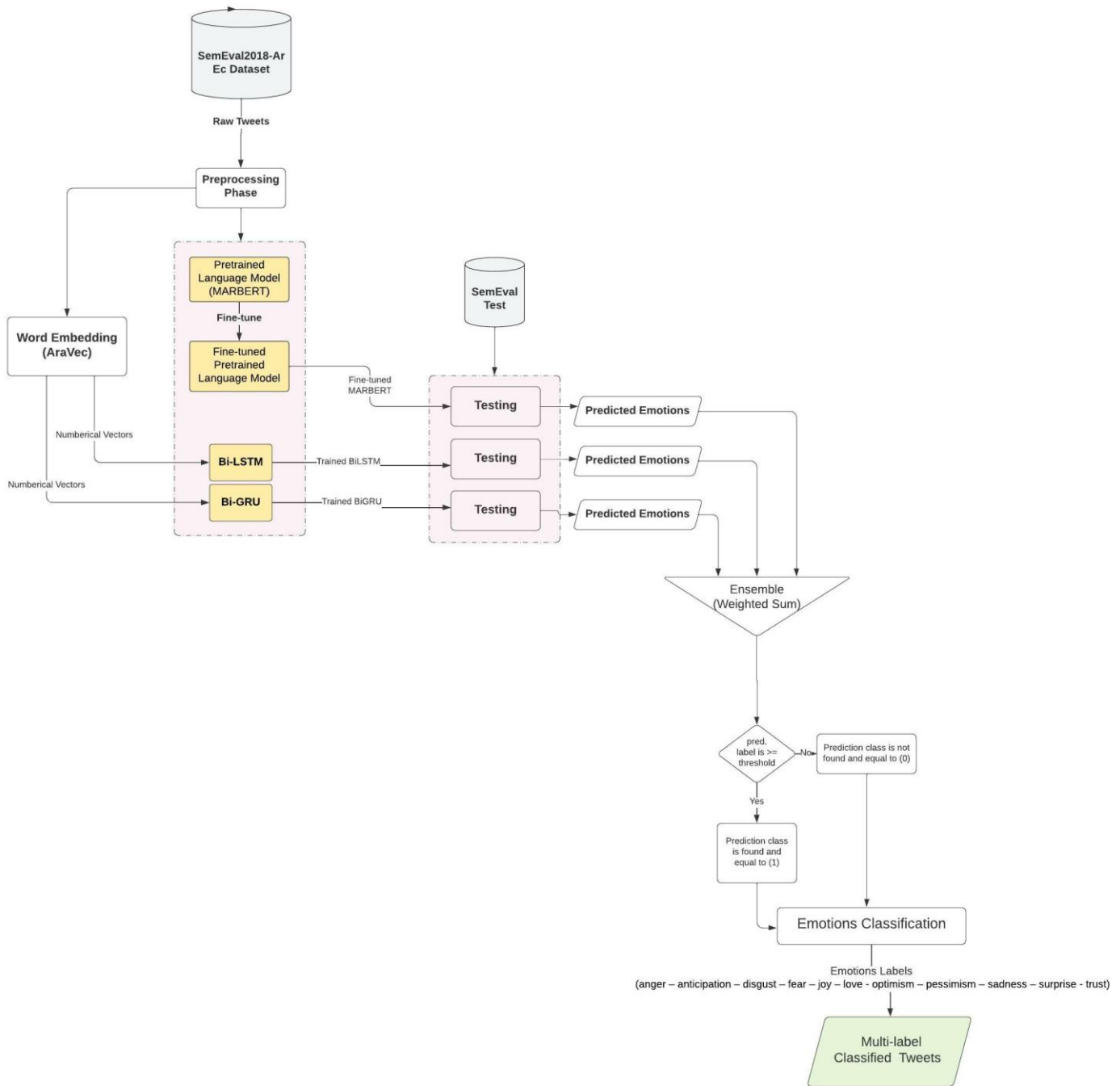


Fig. 2. The Proposed Ensemble Model.

categories (1) Sentiment Analysis SA (2) Named Entity Recognition NER (3) Dialect Identification DI (4) Topic Classification TC (5) Social Meaning SM like (emotion, irony, sarcasm, ...). Because MARBERT transformer was used before in the emotion detection task, it has an emotion-related contextual understanding experience. In this paper, a fine-tuned MARBERT on the SemEval-2018-Ec-Ar task has been proposed.

2) *Bi-LSTM model*: Bidirectional Long Short Term Memory Model or Bi-LSTM is an extension of the normal LSTM introduced by Hochreiter & Schmidhuber in 1997 [21]. LSTM was developed to avoid the short-term dependency problem as it can remember information for long periods, unlike traditional RNNs.

The core component in any LSTM cell is called “cell state”, which maintains information from previous time steps. Addition or deletion to the cell state is controlled by three main gates (forget gate, input gate, and output gate). The input of an LSTM cell is a combination of the input from the current time step and the previous hidden state. This combination outputs a numerical vector whose values are squished between 0 and 1 after applying a sigmoid function as shown in (1). Values closer to 0 will be forgotten while values closer to 1 will be kept and this is called the forget gate.

$$F_t = s (W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

The same combination of input will be copied to be an input for two different activation functions (sigmoid and tanh) which are the main components of the input layer. The output of the sigmoid function is a vector whose values are squished between 0 and 1 by using (2) while the output of the tanh function is a vector whose values are squished between -1 and 1 by using (3). A pointwise multiplication is conducted between the output of these two activation functions which outputs a candidate cell state represented in a vector C_t after filtering non-important information using the sigmoid function.

$$i_t = s (W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$C_t = \tanh (W_c \cdot [h_{t-1}, x_t] + b_c) \quad (3)$$

To calculate the updated cell state, (4) is used which represents a pointwise addition between two parts, the first part is the result of a pointwise multiplication between (the previous cell state and the output of forget gate) while the second part is the result of input gate.

$$C_t = f_t * C_{t-1} + i_t * C_t \quad (4)$$

The final step is to determine the new hidden state, and this is the output gate. To calculate the new hidden state h_t , (6) is used which includes two main parts, the first one is the output of a sigmoid activation function O_t , (5) that accepts a combined input from both the previous hidden state and the current input while the second part is the output of a tanh activation function whose input is the newly updated cell state C_t . As a result, the final output h_t of the LSTM cell will be filtered values from the cell state C_t .

$$O_t = s (W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = O_t * \tanh (C_t) \quad (6)$$

Bi-LSTM is also a type of deep learning model that deals with sequential data. It is an extension of LSTM, and it accepts input data from both directions one from a forward direction and the other from a backward direction. Working in both directions can increase the contextual understanding of the user-generated text. In the following example, the word “احمد” in the first sentence is a noun (a person whose name is Ahmed) but in the second one, it is a verb (means thank). The model can understand the context by working in both directions to decide the meaning of each word in the context based on the next and previous words. In this way, Bi-LSTM can help more in a deep understanding of the user-generated text to get the emotional insights of the tweeter.

اعلان مجدي يعقوب للقلب ❤️ دنيا سمير غانم واحمد حلمي 🐱 بجد
ابدعته الاعلان حلو اووي

دائماً ارفع راسي واحمد ربنا أنى من شعب مصر العظيم.

3) *Bi-GRU model*: GRU or Gated Recurrent Unit is a newer version of the LSTM neural network, and it was introduced by [22]. Unlike LSTM, GRU has fewer steps because it has only two gates (update and reset gates). Also, it has no cell state and the role of maintaining information lies in the hidden state. Update gate acts like forget and update gates in LSTM cell i.e., it decides what information to maintain and what to drop. In most cases, the results of Bi-GRU are slightly faster and better than Bi-LSTM.

D. Testing Layer

After finishing the processing layer, a fine-tuned version of MARBERT is available besides a trained version of Bi-LSTM and Bi-GRU. The SemEval-2018-Ar-Ec test dataset is used to test and evaluate the models. The output of the testing phase is three prediction files ready to enter the ensemble layer.

E. Ensemble Layer

After each model is tested using the SemEval-2018-EC-Ar-test dataset, three prediction files are generated, combined, and processed using a weighted sum equation that balances contextual understanding according to the performance of each model.

F. Classification Layer

Fraction results are generated from the ensemble layer. To get correct values, a certain threshold had been used to determine which values are one “1” meaning class label is found or zero “0” meaning class label is not found.

IV. EMPIRICAL RESULTS AND DISCUSSION

In this section, the results of the proposed ensemble approach are discussed.

A. Dataset

The Arabic dataset has eleven class labels (anger – anticipation – disgust – fear – joy – love – optimism – pessimism – sadness – surprise – trust). For every class label in a tweet there is one of the two binary classification numbers

(zero or one) indicating the feature is found or not. “Zero” means that the emotion is not found while “One” means that it is found. As shown in Table II every tweet is classified into (zero or one) across one or more classes which represent the emotional state of the tweeter. Tweets available in the SemEval2018-Ar-Ec dataset [9] were collected using Twitter API³ and they focused in their searching queries on the tweets related to some emotional words also they have used Best-Worst Scaling (BWS) to determine the annotation reliability. SemEval-2018 Dataset is available for free download from the official site of competition⁴. The dataset is divided into three main files (train, development, and test). Tweets’ count in each file is shown in Table III.

TABLE II. SEMEVAL-2018-AR-EC DATASET DESCRIPTION


Multilabel	Tweet
['anger', 'anticipation', 'disgust', 'fear', 'joy', 'love', 'optimism', 'pessimism', 'sadness', 'surprise', 'trust']	
[0-0-0-1-0-0-0-0-0-0-0-0-0-0-0-0] [Fear]	مومعقول اللي قاعد بصير فيني هالايام يارب ماينتابني شعور الخوف والتوتر اللهم التركيز وأعلى الدرجات
[1-0-0-0-0-0-0-0-0-1-0-0-0-0-0-0] [Anger, Sadness]	احتاج افرغ غضبي على احد بس محد له ذنب فلذلك اتطرق للانعزال
[0-0-0-0-1-1-1-0-0-0-0-0-0-0-0-0] [Joy, Love, Optimism]	كل عام وانت بخير وعيد سعيد وحياتة مليئة بالافراح والمسرات ان شاء الله
[0-0-0-0-1-0-1-0-0-1-0-0-1-0-0-0] [Joy, Optimism, Surprise]	وف لحظه واحده .. تتدخل ارادة ربنا و تحل كل حاجه .. الصبر ! 

TABLE III. NUMBER OF TWEETS IN DATASET FILES

File Name	Number of Arabic tweets
Train	2,278
Development	585
Test (gold labels)	1,518

B. Tools

This work has been implemented on a cloud-based environment “Google Colab”⁵ owned by Google. Colab offers three different plans (Free, Colab Pro, and Colab Pro+) that have differences in RAM, GPUs, storage capacities, and other features. The free plan that provides [12.69 GB of RAM, Python3 Google Compute Engine Backend (GPU), 78.19 GB for Disk Storage] has been utilized. Libraries from “Huggingface”⁶ for working with transformers were utilized. Also, the “simple transformers” library was used for implementing the transformer model.

C. Evaluation Metrics

For the evaluation⁷ of the proposed ensemble model, different evaluation metrics were utilized:

$$\text{Jaccard Accuracy} = \frac{1}{|T|} \sum_{t \in T} \frac{|G_t \cap P_t|}{|G_t \cup P_t|} \quad (7)$$

$$\text{Micro-P} = \frac{\sum_{e \in E} \text{number of tweets correctly assigned to emotion class } e}{\sum_{e \in E} \text{number of tweets assigned to emotion class } e} \quad (8)$$

$$\text{Micro-R} = \frac{\sum_{e \in E} \text{number of tweets correctly assigned to emotion class } e}{\sum_{e \in E} \text{number of tweets in emotion class } e} \quad (9)$$

$$\text{Micro-avg F} = \frac{2 \times \text{Micro-P} \times \text{Micro-R}}{\text{Micro-P} + \text{Micro-R}} \quad (10)$$

$$\text{Precision (P}_e) = \frac{\text{number of tweets correctly assigned to emotion class } e}{\text{number of tweets assigned to emotion class } e} \quad (11)$$

$$\text{Recall (R}_e) = \frac{\text{number of tweets correctly assigned to emotion class } e}{\text{number of tweets in emotion class } e} \quad (12)$$

$$F_e = \frac{2 \times P_e \times R_e}{P_e + R_e} \quad (13)$$

$$\text{Macro-avg F} = \frac{1}{|E|} \sum_{e \in E} F_e \quad (14)$$

The values of True Positives (TP) and True Negatives (TN) are the correct predictions of the classifier while False (FP) and False Negatives (FN) are the mis-predicted values. And the target is to minimize FP and FN.

D. Choosing Best Word Embedding Model

Results of Bi-GRU and Bi-LSTM have been tracked when using two different word embedding models Fasttext⁸ and Aravec⁹. It was found that when applying Aravec, better results are achieved than Fasttext. A comparison between Aravec and Fasttext results is shown in Table IV.

E. Models

1) *Bi-LSTM deep learning model*: The experiments applied using the BiLSTM model were made after defining the parameters shown in Table V.

TABLE IV. A COMPARISON BETWEEN THE RESULTS OF BiGRU AND BiLSTM WHEN USING FASTTEXT AND ARAVEC

Algorithm	Evaluation Metrics				
	Jaccard Score	Precision	Recall	F1 Score	
				Macro	Micro
BiGRU Fasttext/Aravec	0.472 / 0.498	0.434 / 0.599	0.532 / 0.543	0.477 / 0.503	0.642 / 0.664
BiLSTM Fasttext/Aravec	0.455 / 0.485	0.417 / 0.522	0.549 / 0.559	0.469 / 0.509	0.624 / 0.653

³ <https://developer.twitter.com/en/docs/twitter-api>

⁴ <https://competitions.codalab.org/competitions/17751>

⁵ <https://colab.research.google.com/>

⁶ <https://huggingface.co/>

⁷ https://competitions.codalab.org/competitions/17751#learn_the_details-evaluation

⁸ <https://fasttext.cc/>

⁹ <https://github.com/bakriono/aravec>

TABLE V. BI-LSTM MODEL PARAMETERS

Parameter	Value
Cell type	LSTM
Number of cells	2
bidirectional	True
Cell units	256
Max timesteps	64
Batch size	128
Embedding dimensions	300
Learning Rate (LR)	3e-4
No. of epochs	20
Number of classes	11

During the training and validation phases, the model monitors and saves the best checkpoints for different validation metrics like accuracy, precision, recall, and loss. The weights of each best checkpoint have been loaded and made our predictions using the test dataset.

As shown in Table VI, the best recall checkpoint achieved better results than other checkpoints with 0.485, 0.522, 0.559, 0.509, and 0.653 for Jaccard score, precision, recall, and F1 macro & micro scores respectively. Fig. 3, 4, and 5 show the relationship between training and validation for loss, accuracy, and recall at each epoch.

TABLE VI. BEST RESULTS OF BI-LSTM MODEL

Best Metrics	Evaluation Metrics				
	Jaccard Score	Precision	Recall	F1 Score	
				Macro	Micro
BiLSTM_best_precision	0.319	0.210	0.384	0.257	0.483
BiLSTM_best_loss	0.481	0.503	0.512	0.480	0.649
BiLSTM_best_accuracy	0.415	0.400	0.445	0.393	0.586
BiLSTM_best_recall	0.485	0.522	0.559	0.509	0.653

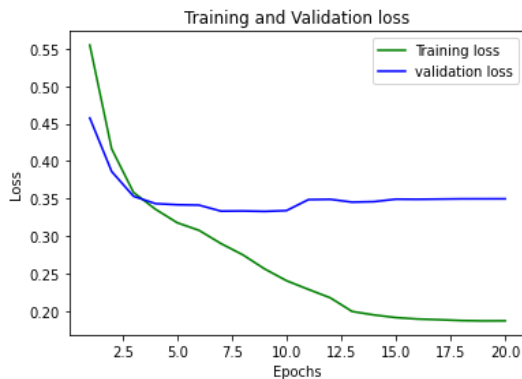


Fig. 3. BiLSTM Training and Validation Loss.

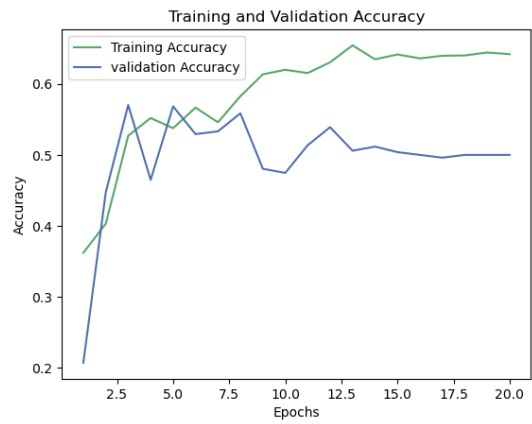


Fig. 4. BiLSTM Training and Validation Accuracy.

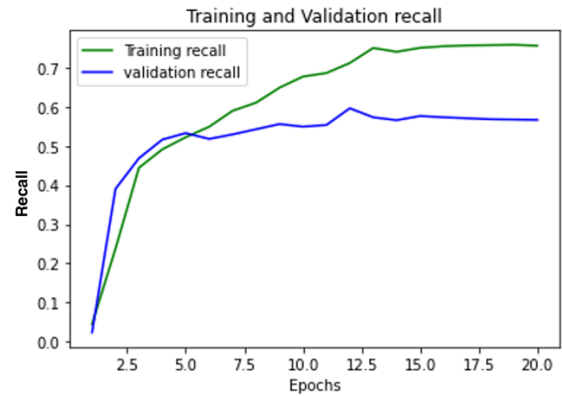


Fig. 5. BiLSTM Training and Validation Recall.

2) *Bi-GRU deep learning model*: The same work was done using Bi-GRU model with the parameters shown Table VII and the results of best checkpoints were compared in Table VIII.

As shown in the comparison, the best recall checkpoint achieved the best prediction results compared to other checkpoints with 0.498, 0.599, 0.503, 0.664 for Jaccard accuracy, precision, macro & micro F1 score, respectively. Fig. 6, 7, and 8 show the relationship between training and validation for loss, accuracy, and recall at each epoch.

TABLE VII. BI-GRU MODEL PARAMETERS

Parameter	Value
Cell type	GRU
Number of cells	2
bidirectional	True
Cell units	256
Max timesteps	64
Batch size	128
Embedding dimensions	300
Learning Rate (LR)	3e-4
No. of epochs	20
Number of classes	11

TABLE VIII. BEST RESULTS OF BI-GRU MODEL

Best Metrics	Evaluation Metrics				
	Jaccard Score	Precision	Recall	F1 Score	
				Macro	Micro
BiGRU_best_precision	0.267	0.248	0.269	0.223	0.422
BiGRU_best_loss	0.498	0.552	0.545	0.500	0.664
BiGRU_best_accuracy	0.431	0.395	0.475	0.426	0.602
BiGRU_best_recall	0.498	0.599	0.543	0.503	0.664

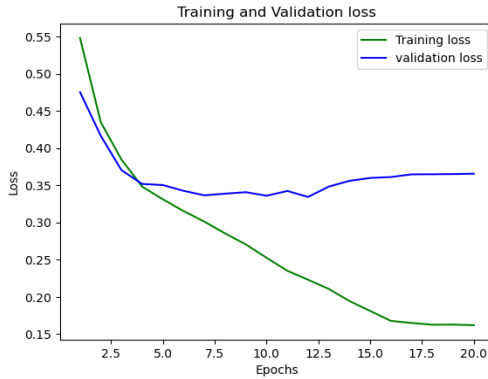


Fig. 6. Bi-GRU Training and Validation Loss.

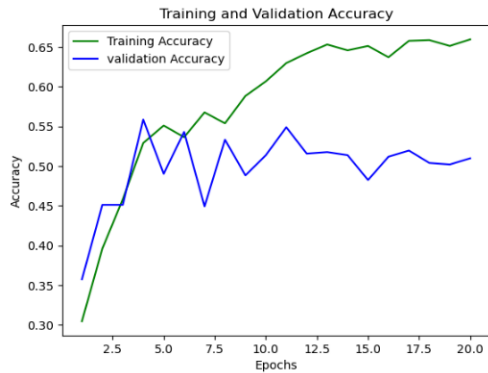


Fig. 7. Bi-GRU Training and Validation Accuracy.

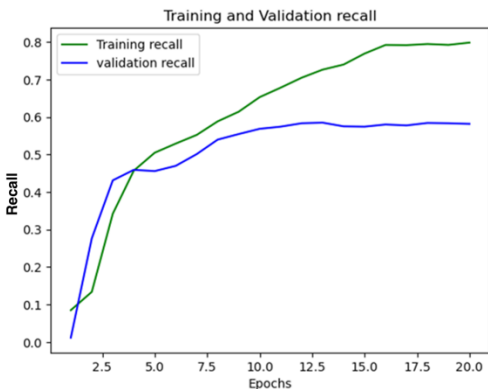


Fig. 8. Bi-GRU Training and Validation Recall.

3) *MARBERT deep learning model*: A pretrained language model “MARBERT” based on BERT was utilized. In [13] authors mentioned that the MARBERT transformer outperforms the recently used ARABERT transformer presented by [4] through their experiments using different datasets. They did not use SemEval-2018 E-c dataset for the emotion classification task. A comparison between the results of both ARABERT and MARBERT against the SemEval-2018-Ar-Ec dataset has been made and the results showed that MARBERT outperforms ARABERT by 4%, 2.7%, 4.2%, and 3.5% regarding Jaccard accuracy, recall, F1 macro, and F1 micro score respectively as shown in Table IX.

TABLE IX. COMPARISON BETWEEN MARBERT AND ARABERT AGAINST SEMEVAL-2018-EC-AR DATASET

Model	Evaluation Metrics				
	Jaccard Score	Precision	Recall	F1 Score	
				Macro	Micro
ARABERT	0.489	0.545	0.532	0.487	0.657
MARBERT	0.529	0.542	0.559	0.529	0.692

In fact, the outperformance is because MARBERT was pretrained on different tasks one of them is related to emotion recognition and it was pretrained on both MSA and Arabic dialects unlike ARABERT as mentioned in Section III.

4) *Ensemble model*: Because each model in the proposed ensemble model has its contextual understanding mechanism, there is a margin difference between the results of MARBERT transformer and the other two models. As a result, a weighted sum equation has been used to ensemble the results of all models i.e., predictions of each model are multiplied by weights according to the percentage of their understanding of the context as shown in (15). If the prediction of a single label is greater than or equal to a certain threshold, the predicted label will equal to “one” or the emotion is found otherwise the emotion is not found, and the predicted label will equal to “zero”. To determine the best threshold and weights a manual grid search was made and gave the best threshold equals 0.34 and the best weights are $w_1 = 0.72$, $w_2 = 0.1$, and $w_3 = 0.18$ for MARBERT, BiLSTM, and BiGRU respectively. The idea of manual grid search is that different weight values ranging from 0.01 to 1 are tested with different thresholds ranging from 0.01 to 1 to find the best result for the proposed ensemble model.

$$P^{total} = (p_1 * w_1 + p_2 * w_2 + p_3 * w_3) \quad (15)$$

where p_1 , p_2 , and p_3 are the predictions of MARBERT transformer, BiLSTM, and BiGRU respectively. P^{total} is the total result of the ensemble model after combining the results of the three models.

A comparison between the results of the closely related studies and the proposed ensemble model is shown in table X where it was found that the ensemble model has achieved the best results regarding Jaccard accuracy, and F1 macro score

with an improvement ranging from 0.2% to 4.2% in accuracy, and from 5.3% to 23.3% in macro F1 score.

Table XI presents a comparison between the proposed ensemble model and each of the separate models that constitute the ensemble model. The table indicates that the best

performance is for the MARBERT transformer model, while the ensemble approach combining all models still has a better effect on the overall performance compared to other models individually with an accuracy of 0.540 and a macro F1 score of 0.701.

TABLE X. CLOSELY RELATED WORKS AND THEIR RESULTS

Reference	Publication	Year	Problem	Dataset	Methods	Performance Measures
A context integrated model for multi-label emotion detection [16]	Elsevier	2018	Arabic Multilabel Emotions Classification	SemEval-2018 task 1-Ec-Ar	C-GRU (Context-aware GRU)	Mic F1: 0.495 Macro F1: 0.648 Jaccard Acc: 0.532
Hybrid Feature Model for Emotion Recognition in Arabic Text [17]	IEEE Access	2020	Arabic Multilabel Emotions Classification	SemEval-2018 task 1-Ec-Ar	HEF + DF Hybrid of human-engineered feature-based model + deep feature-based (DF) model	Micro F1: 0.631 Macro F1: 0.502 Jaccard Acc: 0.512
Pre-Training BERT on Arabic Tweets: Practical Considerations [14]	arXiv	2021	Arabic Multilabel Emotions Classification	SemEval-2018 task 1-Ec-Ar	QARiB Model	Macro F1: 0.468
Combining Context-aware Embeddings and an Attentional Deep Learning Model for Arabic Affect Analysis on Twitter [2]	IEEE Access	2021	Arabic Multilabel Emotions Classification	SemEval-2018 task 1-Ec-Ar	AraBERT word embeddings, attention-based LSTM and BiLSTM	Accuracy: 0.538
Deep Learning for emotion analysis in Arabic tweets [12]	Journal of big data	2021	Arabic Multilabel Emotions Classification	SemEval-2018 task 1-Ec-Ar	Bi-LSTM AraVec / CBOW	Micro F1: 0.61 Precision: 0.695 Recall: 0.551 Jaccard Acc: 0.498
Proposed Ensemble Model	-	2022	Arabic Multilabel Emotions Classification	SemEval-2018 task 1-Ec-Ar	MARBERT, Bi-LSTM, Bi-GRU	Accuracy: 0.540 Macro F1 Score: 0.701 Precision: 0.634 Recall: 0.550 Micro F1 Score: 0.527

TABLE XI. COMPARISON BETWEEN MARBERT, Bi-LSTM, Bi-GRU AND OUR PROPOSED ENSEMBLE MODEL

Model	Evaluation Metrics				
	Jaccard Score	Precision	Recall	F1 Score	
				Micro	Macro
MARBERT	0.529	0.542	0.559	0.529	0.692
Bi-LSTM	0.485	0.522	0.559	0.509	0.653
Bi-GRU	0.498	0.599	0.543	0.503	0.664
Proposed Ensemble Model	0.540	0.634	0.550	0.527	0.701

V. CONCLUSION AND FUTURE WORK

ED can help communities in different domains, especially for social networks which have many signed users sharing their feelings and thoughts. Understanding the context of user-generated Arabic text still has a lot of challenges because of the language complexity, the limited Arabic resources, and different available Arabic dialects. Recent research studies have used pre-trained language models to overcome the issue of limited resources. In this paper, a pretrained language model (MARBERT) is fine-tuned using the SemEval-2018-task1-ArEc dataset that was published in SemEval-2018-task1: Affect in Tweets to perform a multilabel classification task. Three state-of-the-art models (BiLSTM, BiGRU, and MARBERT) were ensembled and compared to recently published studies. The experimental results showed that the proposed ensemble model outperforms the best existing related work with an improvement ranging from 0.2% to 4.2% in accuracy, and from 5.3% to 23.3% in macro F1 score.

Also, it was noticed that the SemEval-2018 dataset we are using in this paper is not balanced. Three classes (anticipation, surprise, and trust) have low instances in the dataset making the dataset imbalanced. So, different data augmentation and oversampling techniques can be applied to solve this issue in future studies.

REFERENCES

- [1] O. AlZoubi, S. K. Tawalbeh, and M. AL-Smadi, "Affect detection from arabic tweets using ensemble and deep learning techniques," *Journal of King Saud University - Computer and Information Sciences*, Oct. 2020, doi: 10.1016/j.jksuci.2020.09.013.
- [2] H. Elfaik and E. H. Nfaoui, "Combining Context-Aware Embeddings and an Attentional Deep Learning Model for Arabic Affect Analysis on Twitter," *IEEE Access*, vol. 9, pp. 111214–111230, 2021, doi: 10.1109/ACCESS.2021.3102087.
- [3] Z. Liu, A. Xu, Y. Guo, J. U. Mahmud, H. Liu, and R. Akkiraju, "Seemo," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Apr. 2018, pp. 1–12. doi: 10.1145/3173574.3173938.
- [4] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," Feb. 2020, [Online]. Available: <http://arxiv.org/abs/2003.00104>.
- [5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018, [Online]. Available: <http://arxiv.org/abs/1810.04805>.
- [6] N. Alswaidan and M. E. B. Menai, "A survey of state-of-the-art approaches for emotion recognition in text," *Knowledge and Information Systems*, vol. 62, no. 8, pp. 2937–2987, Aug. 2020, doi: 10.1007/s10115-020-01449-0.
- [7] F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah, "Text-based emotion detection: Advances, challenges, and opportunities," *Engineering Reports*, vol. 2, no. 7, Jul. 2020, doi: 10.1002/eng2.12189.
- [8] D. William and D. Suhartono, "Text-based Depression Detection on Social Media Posts: A Systematic Literature Review," *Procedia Computer Science*, vol. 179, pp. 582–589, 2021, doi: 10.1016/j.procs.2021.01.043.
- [9] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 Task 1: Affect in Tweets," in *Proceedings of The 12th International Workshop on Semantic Evaluation*, 2018, pp. 1–17. doi: 10.18653/v1/S18-1001.
- [10] G. Badaro *et al.*, "EMA at SemEval-2018 Task 1: Emotion Mining for Arabic," in *Proceedings of The 12th International Workshop on Semantic Evaluation*, 2018, pp. 236–244. doi: 10.18653/v1/S18-1036.
- [11] M. Baali and N. Ghneim, "Emotion analysis of Arabic tweets using deep learning approach," *Journal of Big Data*, vol. 6, no. 1, p. 89, Dec. 2019, doi: 10.1186/s40537-019-0252-x.
- [12] E. A. H. Khalil, E. M. F. E. Houbay, and H. K. Mohamed, "Deep learning for emotion analysis in Arabic tweets," *Journal of Big Data*, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-021-00523-w.
- [13] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic," Dec. 2020, [Online]. Available: <http://arxiv.org/abs/2101.01785>.
- [14] A. Abdelali, S. Hassan, H. Mubarak, K. Darwish, and Y. Samih, "Pre-Training BERT on Arabic Tweets: Practical Considerations," Feb. 2021, [Online]. Available: <http://arxiv.org/abs/2102.10684>.
- [15] N. Al-Twairesh, "The Evolution of Language Models Applied to Emotion Analysis of Arabic Tweets," *Information*, vol. 12, no. 2, p. 84, Feb. 2021, doi: 10.3390/info12020084.
- [16] A. E. Samy, S. R. El-Beltagy, and E. Hassanien, "A Context Integrated Model for Multi-label Emotion Detection," *Procedia Computer Science*, vol. 142, pp. 61–71, 2018, doi: 10.1016/j.procs.2018.10.461.
- [17] N. Alswaidan and M. E. B. Menai, "Hybrid Feature Model for Emotion Recognition in Arabic Text," *IEEE Access*, vol. 8, pp. 37843–37854, 2020, doi: 10.1109/ACCESS.2020.2975906.
- [18] B. Talafha *et al.*, "Multi-Dialect Arabic BERT for Country-Level Dialect Identification," Jul. 2020, [Online]. Available: <http://arxiv.org/abs/2007.05612>.
- [19] 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.
- [20] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, "AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP," *Procedia Computer Science*, vol. 117, pp. 256–265, 2017, doi: 10.1016/j.procs.2017.10.117.
- [21] A. Vaswani *et al.*, "Attention Is All You Need," Jun. 2017, [Online]. Available: <http://arxiv.org/abs/1706.03762>.
- [22] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
- [23] K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," Jun. 2014, [Online]. Available: <http://arxiv.org/abs/1406.1078>.