University of Baghdad
College of Engineering
JE
JOURNAL OF ENGINEERING

Journal of Engineering

journal homepage: www.joe.uobaghdad.edu.iq

Volume 29 Number 9 September 2023

Iraqi Sentiment and Emotion Analysis Using Deep Learning

Anwar Abdul-Razzaq Alfarhany [1,*], Nada A. Z. Abdullah[2]

Department. of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq
Anwar.Abdulrazzaq1201a@sc.uobaghdad.edu.iq[1], Nada.abdullah@sc.uobaghdad.edu.iq[2]

ABSTRACT

Analyzing sentiment and emotions in Arabic texts on social networking sites has gained wide interest from researchers. It has been an active research topic in recent years due to its importance in analyzing reviewers' opinions. The Iraqi dialect is one of the Arabic dialects used in social networking sites, characterized by its complexity and, therefore, the difficulty of analyzing sentiment. This work presents a hybrid deep learning model consisting of a Convolution Neural Network (CNN) and the Gated Recurrent Units (GRU) to analyze sentiment and emotions in Iraqi texts. Three Iraqi datasets (Iraqi Arab Emotions Data Set (IAEDS), Annotated Corpus of Mesopotamian-Iraqi Dialect (ACMID), and Iraqi Arabic Dataset (IAD)) collected from Facebook are used to evaluate the model. Experiments showed that the model obtained good results, as the accuracy of the model was 91.1, 92.4, and 92.5% for IADS, ACMID, and IAD, respectively. The results of the model outperformed previous works for all datasets.

Keywords: Emotion analysis, Sentiment analysis, CNN, GRU, Iraqi dialect.

*Corresponding author
Peer review under the responsibility of University of Baghdad.
https://doi.org/10.31026/j.eng.2023.09.11

Article received: 17/02/2023
Article accepted: 15/04/2023
Article published: 01/09/2023

تحليل المشاعر والعواطف في النصوص العراقية باستخدام التعلم العميق

انوار عبدالرزاق الفرحاني[1]،*، ندا عبدالزهرة عبدالله[2]

قسم علوم الحاسوب، كلية العلوم، جامعة بغداد، بغداد، العراق

الخلاصة

حظي تحليل المشاعر والعواطف في النصوص العربية على مواقع التواصل الاجتماعي باهتمام واسع من الباحثين، حيث أصبح موضوع بحث نشط في السنوات الأخيرة، لما له من أهمية في تحليل آراء المراجعين في مختلف الموضوعات. تعتبر اللهجة العراقية من اللهجات العربية المستخدمة في مواقع التواصل الاجتماعي، وتتميز بتعقيدها، وبالتالي صعوبة تحليل المشاعر فيها. في هذه العمل، تم تقديم نموذج هجين للتعلم العميق يتكون من CNN وGRU لتحليل المشاعر والعواطف في النصوص العراقية على وجه التحديد. تم استخدام ثلاث مجموعات بيانات عراقية، وهي IADS وACMID وIAD، تم جمعها من Facebook لتقييم النموذج. أظهرت التجارب أن النموذج حصل على نتائج جيدة حيث كانت دقة النموذج 91.1٪ و92.4٪ و92.5٪ لكل من IADS وACMID وIAD على التوالي. تفوقت نتائج النموذج على الأعمال السابقة لجميع مجموعات البيانات.

الكلمات المفتاحية: التعلم العميق، تحليل المشاعر، اللهجة العراقية.

1. INTRODUCTION

With the development of the Internet and its availability almost everywhere, Sentiment Analysis (SA) has become one of the most interesting research areas in natural language processing. Users can now express their opinions and consult each other about services, events, people, institutions, products, and goods before purchasing them in various ways, including product reviews, opinions expressed, and customer feedback. This data type has become the focus of researchers' attention. Most decision-makers or commercial companies try to understand the feelings of citizens and customers because of its benefit in improving their products or services to attract customers **(Oueslati et al., 2020)**.

SA, or Opinion Mining (OM), relies heavily on polarity categorization. The polarity of a sentence, paragraph, or even a single word refers to the overarching feeling it expresses, which can be a positive or negative feeling, what is known as a binary classification of feelings **(Abu Kwaik et al., 2019)**. In certain circumstances, sentiment analysis may fall short of capturing the reviewer's or writer's feelings. Therefore, most studies have tended to emotion analysis to understand the underlying emotions within these opinions and reviews. Emotion analysis is considered an element of sentiment analysis due to examining the writer's inner feelings, which recognize particular emotions rather than positive and negative sentiments **(Sailunaz and Alhajj, 2019)**. Both Ekman model **(Ekman, 1992)** and the Plutchik model **(Plutchik, 1980; Plutchik, 1994)** are examples of emotional representation models. Ekman displays a wide range of emotions, including Anger, sadness, happiness, fear, disgust, and surprise. On the other hand, the Plutchik is made up of Ekman's six emotions plus two labels, namely, trust and anticipation.

Studies on the topic of sentiment analysis or emotion analysis in the Arabic language were few compared to those in the English language, particularly those in the dialects of the Arabic language, which were rare because most of the tools were focused on the standard Arabic

language. However, the majority of Internet users write in their native dialects that they use in their daily lives **(Medhat et al., 2014).**

This study used a hybrid deep learning model comprised of two different neural network types, Convolution Neural Network (CNN) and Gated Recurrent Units (GRU), to analyze sentiment and emotion in the Iraqi text. The model consists of three stages. The first stage is the pre-processing stage, which was appropriate for the Iraqi dialect, the stage of word embedding, and then the classification stage. The feature vectors obtained from the embedding stage are passed to the CNN layer, then to two layers of the GRU, and finally to the dense layer to classify the text.

1.1 Arabic Sentiment Analysis Related Work

(Alayba et al., 2018b) proposed an Arabic sentiment analysis model for the dataset they had collected on subjects related to health services topics. They employed different Arabic corpus to construct Word2Vec models, which had been used to train CNN and lexicon models to get around the problem of training on a limited dataset. The proposed model's accuracy outperformed the previous one, reaching 92% and 95% for Main-AHS and Sub-AHS, respectively. The effectiveness of combining CNN and Long Shot Term Memory (LSTM) models was also examined by **(Alayba et al., 2018a)** to enhance the sentiment analysis of the same Arabic datasets. The usefulness of using several levels of word embeddings, character level, word level, and ch5gram level, was also explored due to the challenges of morphology and spelling in Arabic. where the CNN layer receives the input first, followed by the LSTM layer. The output layer makes the final prediction using a sigmoid function. Tests showed that the Main-AHS and Sub-AHS datasets' sentiment classification accuracy increased to 0.9424 and 0.9568, respectively. A CNN and a bi-directional LSTM (BILSTM) ensemble model were combined by **(Heikal et al., 2018)** to make textual sentiment predictions in Arabic tweets utilizing the Arabic Sentiment Tweets Dataset (ASTD). BILSTM, CNN, and the ensemble model were used in their separate studies. The experiments' results demonstrated that the ensemble model had a high F1-score and accuracy compared to the others, reaching 64.46% and 65.05%, respectively. A deep hybrid model was created by **(Wint et al., 2018)** by combining two CNN and BILSTM models (H2CBi). They used data from (Yelp, Twitter I, Twitter II, Facebook, and Form Spring. me) and two other types of social network services (SNS), including review data from Yelp, Movie Review, and Amazon. Data features, such as whether they were positive, negative, bullied, or not, were extracted using a CNN model, and a BILSTM was utilized to represent sentences. They utilized word2vec, fastText, and Glove pre-trained word embedding for the word representation. Three of six models of H2CBi outperformed the baseline result on six of the seven datasets according to the experiments performed on them. **(Mohammed and Kora, 2019)** offered three models to analyze the polarity in a corpus of forty thousand texts written in MSA and Egyptian dialects: CNN, LSTM, and recurrent convolution neural network (RCNN). According to the tests they performed to compare the three models, LSTM outperformed the other two models, with an accuracy rate of 81.31%. Additionally, using the corpus augmentation technique increased accuracy to 88.05%. **(Alnawas and Arici, 2019)** used four binary classifiers to detect feelings in Iraqi texts (Decision Tree, Logistic Regression (LR), Naive Bayes, And Support Vector Machine (SVM)) with varying parameter values (dimensions, window size, and negative samples). To create a word embedding model, they gathered a significant body of previous work in the field and trained it using Doc2Vec representations based on the Paragraph and Paragraph Vector (DM-PV) distributed memory model. The

results of the experiments revealed that the logistic regression and the SVM obtained the highest F1-score, 77% and 78%, respectively. Three models of deep learning with various neural network architectures, including CNN, LSTM, and CNN-LSTM combination models, were used by **(Elzayady et al., 2020)** to analyze sentiment on two datasets, Hotels Reviews (HTL) and Book Reviews (LABR). According to the experimental findings, the combined model outperformed CNN and LSTM separately, scoring 85.83% for HTL and 86.88% for LABR. **(Ombabi et al., 2020)** constructed a combination model consisting of a CNN as one layer, LSTM as two layers, and an SVM as a classifier at the final stage to categorize the polarity of the text in various topics. The words were represented as vectors by Fasttext. Multiple experiments' findings demonstrated that the proposed model performed well, with an accuracy of 90.75%. **(Nassif et al., 2021)** provided an in-depth investigation of the performance of shallow learning classifiers and deep learning models for Arabic review sentiment analysis, including CNN, LSTM, GRU, and their hybrids. State-of-the-art models like the araBERT pre-trained model and the transformer architecture are also included in the comparison. Multi-dialect Arabic hotel and book review datasets, some of the biggest freely accessible datasets for Arabic reviews, were used. In terms of classification for binary and multiple labels, the results demonstrated that deep learning exceeded shallow learning. The performance of deep learning models using the default embedding layer was similar, while the implementation of araBERT improved the effectiveness of the transformer model. **(Khabour et al, 2022)** developed a semantic orientation strategy based on a built-in domain ontology and the accessible sentiment lexicon for determining overall polarity from Arabic subjective writings. They used the ontology concepts' levels in the ontology tree and the frequencies with which they appeared in the dataset to extract and weight the semantic domain variables necessary for determining the overall polarity of a textual review. The hotels' domain Arabic dataset was utilized to build the domain ontology and test the proposed method. The f-measure and total accuracy both maximum out at 78.75% and 79.20%, respectively. Based on the results, it seems to be a promising method for application in Arabic sentiment analysis, outperforming previous semantic orientation methods in this area. **(Saleh et al., 2022)** have proposed a methodology for applying a heterogeneous stacking ensemble to enhance the performance of Arabic sentiment analysis. In order to improve the model's performance for predicting Arabic sentiment analysis, it integrates three separate pre-trained Deep Learning models (LSTM, GRU, RNN) with three meta-learners (LR, SVM and Random Forest (RF)). Three benchmark Arabic datasets (the Arabic Sentiment Twitter Corpus (ASTC), ArTwitter, and Arabic Jordanian General Tweets (AJGT)) are used to evaluate the effectiveness of the proposed approach. With an accuracy of 98.08%, 92.24%, and 93.4% for the mentioned data sets, respectively, for the datasets, the results show that the model outperforms other models on each dataset.

1.2 Arabic Emotion Analysis Related Work

(Badaro et al., 2018) presented the first study in Arabic emotion analysis using the SemEval 2018 Task 1: Affect in Arabic Tweets **(Mohammad et al., 2018),** which had five subtasks. The features were extracted using an AraVec pre-trained model **(Soliman et al., 2017).** To classify tweets, use multiple models, such as a CNN, LSTM, Ridge regression, SVM, random forests, or ensemble approaches. After running trials and comparing the results with the other models, it was discovered that the SVM performed best, with an accuracy of 48.9 %. Two sub-models were presented by **(Abdullah et al., 2019)** to identify the sentiment and emotion in Arabic tweets and evaluate their intensity. SemEval-2018 Task 1 **(Mohammad**

et al., 2018) Arabic tweets were used. For the first sub-model, they were used as data inputs in two different ways: in their original Arabic form (ArTweets). Then, they were translated into English (TraTweets) and represented by a 4,908-dimensional vector, where they used a set of semantic features and word and document embeddings to extract features. The input vector is fed to a fully connected neural network with three hidden dense layers. Unlike the second sub-model, where solely ArTweets was used, the AraVec model was used for word embedding, which has a 300-dimensional vector to represent each word in ArTweets. The second sub-model consists of one layer of a CNN and max pooling. One layer of LSTM was added, and two dense layers were added. Finally, both sub-models employed a sigmoid function at the output layer to determine the emotional intensity or sentiment between 0 and 1. For the five sub-tasks of SemEval-2018 Task 1 **(Mohammad et al., 2018; Abdullah and Shaikh, 2018)** created a system to analyze emotions for both English tweets (EngTweets) and Arabic tweets, where the Arabic tweets were supplied as both original Arabic (AraTweets) and translated English tweets (TraTweets). The word was transformed into a feature vector using the AffectiveTweets Weka package, document and word embedding, and the AraVec model for AraTweets. Dense-Network and LSTM-Network, each with one sigmoid neuron at the output layer, were utilized to identify the emotional intensity between 0 and 1. The system exceeded the baseline model for each subtask, which indicated that the performance was good. A deep CNN model was suggested by **(Baali and Ghneim, 2019)** to detect the emotions in the Arabic tweets provided by SemEval for the EI-oc task. The Word2vec model was built using the Genism library and trained using their datasets to convert words into vectors. The network was constructed in four main steps: word, sentence, and document vectorization, followed by classification. There were seven layers total: input (word vectorization), convolution (sentence vectorization), max pooling, flatten, concatenate, dense with ReLU activation (classification), as well as an output using four neurons with softmax (classification). **(Almahdawi and Teahan, 2019)** collected an Iraqi dataset from Facebook based on Ekman's six emotions (happy, anger, sad, fear, disgust, and surprise). They use WEKA's four standard classifiers (ZeroR, J48, Naive Bayes, Multinomial Naive Bayes for text and SMO) and an external classifier called Prediction by Partial Matching (PPM). The results showed that the PPM model was superior to the rest of the models, as it achieved an accuracy of 87.1% and 0.59, 0.63, and 0.61 for recall, precision, and F-measure, respectively. **(Alswaidan and Menai, 2020)** explored three models to detect emotion in three Arabic datasets: SemEval- 2018, IAEDS, and Arabic emotions Twitter dataset (AETD): a Human-Engineered Feature-based (HEF) model, a Deep Feature-based (DF) model, and a hybrid of both models (HEF+DF). The trials' findings demonstrated that the HEF+DF model was superior to the DF and HEF models on all datasets. **(Khalil et al., 2021)** developed a multilabel classification model, which is a multilayer BiLSTM network, and used Arabic tweets from the SemEval 2018 task1 -sub-task E-C to test the model. A 300-dimensional vector was created using the AraVec embedding pre-trained model to represent each word. The tweet embeddings are calculated using the average embedding for all of the words in the tweet. Three BiLSTM layers received input from the tweet's embedded vector. Finally, a dense layer with 11 outputs matching the 11 emotions was used for classification using a sigmoid function. A deep learning ensemble method was suggested by **(Mansy et al., 2022)** to analyze user-generated text from Twitter regarding the emotional insights that reflect different feelings. Three state-of-the-art deep learning models formed the basis for the suggested model. Two models are extensions of the RNN (Bi-LSTM and Bi-GRU), while the third is a MARBERT transformer, a pre-trained language model (PLM) inspired by BERT. Experiment evaluations were conducted using the SemEval-2018-Task1-Ar-Ec dataset, It

was presented at the SemEval-2018 competition as part of a multilabel classification challenge called "Emotion Classification" (EC). MARBERT PLM is comparable to AraBERT, a well-known PLM for working with the Arabic language. According to what was found from the experiments, MARBERT improved the outcomes by 4% in Jaccard accuracy, 2.7% in recall, 4.2% in F1 macro, and 3.5% in F1 micro. Additionally, the suggested ensemble model outperformed the standalone models (Bi-LSTM, Bi-GRU, and MARBERT). Furthermore, it surpasses the most current comparable work by an accuracy margin of 0.2% to 4.2% and a macro F1 score advantage of 5.3% to 23.3%.

2. SENTIMENTAL\EMOTIONAL EXTRACTION METHODOLOGY

2.1 Datasets

Three datasets written in the Iraqi dialect were collected from the social networking site Facebook to test the deep learning models, as follows:

2.1.1 Iraqi Arab Emotions Data Set (IAEDS)

It is the first corpus for emotion recognition written in the Iraqi dialect. It was a collection of posts manually collected from Facebook according to the emotional state of the post that appears at the top of the post, based on Ekman's six emotions (Anger, Surprise, Happiness, Fear, Disgust, and Sadness). The dataset is 1,365 posts, distributed unbalanced over six files, one for each of the six emotions (Fear 148 posts, Disgust 185 posts, Surprise 229 posts, Sad 238 posts, Happy 256 posts, and Anger 309 posts) **(Almahdawi and Teahan, 2019).**

2.1.2 Iraqi Arabic Dataset (IAD)

It is comments collected from Facebook, using the Facepager tool, from Iraqi pages such as the Airways Company, Restaurants, News, Sports, Communications Company, and home appliances company. Three local experts categorized 2,000 comments as 1,000 positive and 1,000 negative. It has been pre-preprocessed by deleting punctuation marks, non-Arabic letters, stop words, short vowels, and diacritical marks (harakat), in addition to normalizing each of the alif, Hamza, and alif maqsoura **(Alnawas and Arici, 2019)**.

2.1.3 Annotated Corpus of Mesopotamian-Iraqi Dialect (ACMID)

It is 5,000 Facebook comments collected from popular Iraqi pages on different topics, namely the University of Baghdad page, Iraqi restaurants, and the show of Melon City. Two Iraqi annotators categorized each comment according to their polarity (positive, negative, neutral, and spam) **(Askar and Sjarif, 2021).**

2.2 Pre-processing

The preprocessing steps were as follows:

- Remove digits, symbols, and special characters, whether Arabic or foreign.
- Remove English, Latin, or any non-Arabic Characters.
- Remove punctuation marks.
- Remove single characters.

- Because emojis are very important for analyzing and recognizing emotions **(Nassr et al., 2020)**, we replaced them with an Arabic word that expresses them, according to the dataset that they found. The rest of the emojis in the form of things, animals, or do not contain expressions have been deleted. **Table 1** shows the replacement process.

Table 1. Emoji Replacement with Arabic Words

Emojis	Emotion	Arabic Word	Sentiment	Arabic Word
😠 🤬 😈 👿 😡, ...	Anger	غضب	Negative	سلبي
😂 🤢 🤮 😑 😐, ...	Disgust	قرف	Negative	سلبي
😨 😰 😱 😫 😟, ...	Fear	خوف	Negative	سلبي
😭 😔 😕 💔 😢, ...	Sad	حزن	Negative	سلبي
😍 🤗 😂 😆 😄, ...	Happy	سعادة	Positive	ايجابي
😳 😮 😲 😦 😶, ...	Surprise	مفاجئة	Positive	ايجابي

- A normalization task was also applied, where some words or letters were normalized as follows:
- Removing Diacritics (Arabic Vowel Marks) "ً ٌ ٍ َ ُ ِ ّ ْ" and Tatweel character "ـ" by using Tashaphyne stemmer (Zerrouki, 2018).
- Transform three Arabic letters Alef forms "أ, إ, آ" into normal Alef "ا."
- Transform the Arab letter Alef Maqsoura "ى" into the Arabic letter Ya "ي."
- Transform the Arabic letter Teh Marbuta "ة" into the Arabic letter Heh "ه."
- For a word that contains repeated letters that are not from the origin of the word but have been added for elongation to express certain emotions, a list has been prepared that includes several words that contain repeated letters from the origin of the word, for example, (ممنوع, حررها, الله, ...), as the repetition does not exceed two letters. The words from the dataset that contain repeated letters are compared with those in the list, and if they are not found, the repetition is deleted.
- Transform any non-Arabic letter similar to the Arabic letter in writing into its corresponding Arabic letter.
- The stemming task was applied using DFSA stemmer **(Abdulhameed, 2020)** to stem the Iraqi dataset, where the prefixes connected to the Iraqi word were deleted. If the word consists of at least 5 letters, prefixes such as ("ال", "و", "د", "ش", "ح", "شد", "بهال", "بهل", "وال", "بل") will be omitted. If the word consists of at least 6 letters, prefixes such as ("ب", "ال") will be omitted.

2.3 Word Embedding

Word embedding is a feature-learning approach where words are mapped to vectors with their semantic relationships and meanings encoded using the contextual hierarchy of words. Comparable words will have the same feature vectors. Feature vectors were created using libraries to build models, using various techniques and vector dimensions, or using pre-trained models.

Since the used datasets are written in the Iraqi dialect, often when pre-trained models are used to convert words into vectors, many words that do not have vectors appear, which are known as Out-of-Vocabulary **(Alswaidan and Menai, 2020)**. An Iraqi Word2Vec model was created by utilizing a Gensim application, as described by **(Word2vec Embeddings)**. The

model was trained on three preprocessed datasets, which all were collected in one dataset and passed to the applications. It was implemented using the Continuous Bag of Words (CBOW) and Skip-Gram (SG) architectures. Two window sizes (50 and 100) were used along with them. Thus, four types of vectors are created, becoming the initialization weights of the embedding layer, which usually start with random weights. The CBOW technique is used to forecast the target or center words from the words surrounding them inside the window's length. On the other hand, the SG technique uses the center word to infer surrounding words.

2.3.1 Deep Learning Model

A hybrid deep learning model was built consisting of one CNN layer followed by two layers of GRU to analyze sentiment and emotions in Iraqi texts, according to the classification of the datasets used, as shown in **Figure 1**.

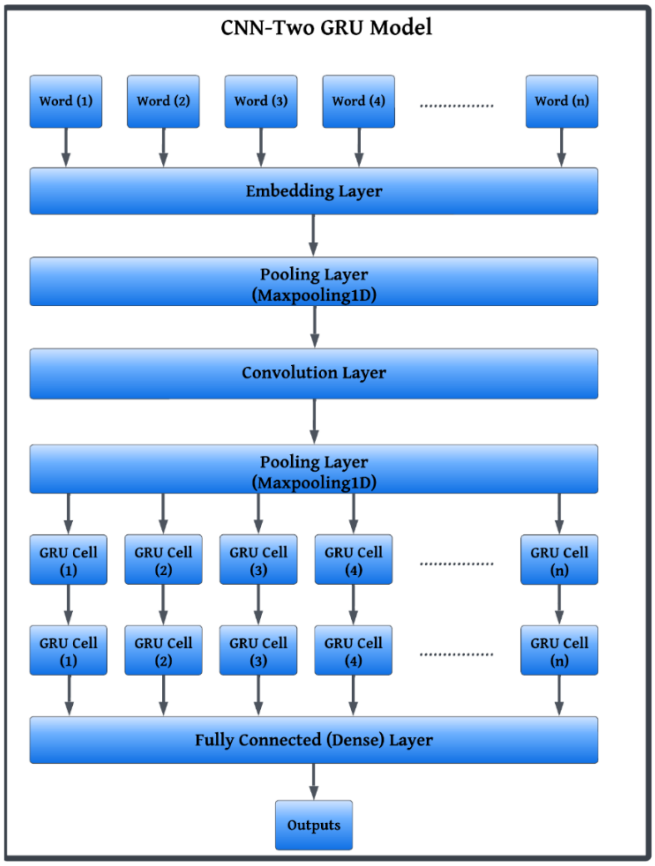

Figure 1. Hybrid CNN-Two GRU Model.

2.3.2 Input Layer

The input layer in the model is an embedding layer that represents each sentence (post or comment) as a row of vectors. Each vector represents a word in the post, where each word is embedded into vectors of different lengths. This layer is a matrix with dimensions $w * v$, where w is the total number of words in the post or the maximum post length, and v is the vector length representing a single word. In cases where a post is less than the allowed maximum, it will be extended by adding zeros until it reaches the permitted length.

2.3.3 Pooling Layer

After the embedding layer is fed to the network, the features of the embedding layer are down-sampled using the pooling layer so that the set of features becomes smaller and the best relationships between features are discovered for classification. A nonlinear down-sampling technique called “Max pooling” is employed, which aids in selecting the words or features that perform the best. Where the down-sampling is applied by computing the maximum activation of predefined subregions within the features set.

2.3.4 Convolution Layer

A fixed-size filter scans a sequence of vectors in the input layer to extract the n-grams of features. 256 filters were used with a size of 5 to extract the 5-gram features of words. To represent multiple features in a post in the feature map, each filter employed the ReLU activation function to identify them. If x is positive, it outputs x; else, it outputs 0.

2.3.5 Pooling Layer

Using the max pooling method, the pooling layer is added again to down-sample the feature map generated by the previous convolution layer. Where the best performance features are selected to be fed to the next layer by computing the maximum activation of predefined subregions within the feature map.

2.3.6 Two GRU Layers

Two layers of GRU are added to the model after the CNN layer one after the other, each having a total number of units of 100, and each layer will output all hidden states for each time step.

2.3.7 Dense Layer

The last layer added converts the input vector from the previous layer into a single output according to how many classes there are in it. The number of classes is two for binary classification and four and six for multiclassification. The activation function in this layer is the sigmoid function with binary classification and the softmax function with multiclassification.
The dropout of the recurrent layer was previously set at 0.3 to minimize the overfitting issue after the second GRU layer. To get the best results, the Adam optimizer, "categorical cross-entropy" and "binary cross-entropy" loss functions, "accuracy" metrics, and early stopping were used.

3. RESULTS AND DISCUSSION

All experiments have been implemented in Python 3.7.14 on the Google Colaboratory platform running on a 12-GB GPU. The following libraries have been used: Keras 2.8.0 under TensorFlow 2.8.2 and Gensim 3.6.0. Three processed datasets were used to test and evaluate

the proposed CNN-GRU model to analyze sentiment and emotions in them. Metrics such as *accuracy, precision, recall, and F1-score* were used to calculate model evaluation for the IAD, and *Jaccard accuracy, macro-averaged precision, macro-averaged recall, and macro-averaged F-score* for IEADS and ACMID, which is calculated in Eq. (1 to11) **(Alswaidan and Menai, 2020):**

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

where TP denotes a True Positive, TN a True Negative, FP a False Positive, and FN a False Negative.

$$F1-score = 2 \cdot \frac{Precision \cdot Recall}{Precision \cdot Recall} \quad (4)$$

$$Jaccard\ accuracy = \frac{1}{|S|}\sum_{s\in S}\frac{|G_s \cap P_s|}{|G_s \cup P_s|} \quad (5)$$

Where S for the collection of sentences, G_s stands for the set of gold labels for sentence*s* and P_s for the set of predicted labels for sentence *s*.

Precision and *recall* are computed for each emotion label *e* on its own, and then the average is obtained for the macro-averaged findings. Hence, each emotion label is given the same weight.

$$precision_e = \frac{TP_e}{TP_e + FP_e} \quad (6)$$

$$recall_e = \frac{TP_e}{TP_e+FN_e} \quad (7)$$

The $F1-score_e$ is the harmonic mean of the two equations given above:

$$F1-score_e = 2 \cdot \frac{precision_e \times recall_e}{(precision_e+recall_e)} \quad (8)$$

The P^{macro} and R^{macro} are calculated as follows:

$$P^{macro} = \frac{1}{|E|}\sum_{e\in E} precision_e \quad (9)$$

$$R^{macro} = \frac{1}{|E|}\sum_{e\in E} recall_e \quad (10)$$

and the F^{macro} is the harmonic mean of the two equations given above:

$$F^{macro} = 2 \cdot \frac{P^{macro} \times R^{macro}}{(P^{macro}+R^{macro})} \quad (11)$$

Since the IADS dataset and the ACMID dataset are unbalanced. The stratified k-fold cross-validation method was chosen to split datasets, with k equals 10. Ensuring that in each split, the distribution of each class would match its distribution in the complete training dataset. The third dataset is well-balanced, but unlike the previous two, it has been split into a training and testing set using the Hold-out cross-validation method, with 80 percent for training and 20 percent for testing. The results of the proposed model for the IAEDS dataset are shown in **Fig. 2**, where the best results were obtained when using the Iraqi model with a technique SG and a vector length equal to 50. Results of the model concerning the ACMID dataset are shown in **Fig. 3**. The best performance of the proposed model was recorded with the technique SG of the Iraqi model with the length of vector 100. Finally, the results of the IAD dataset are shown in **Fig. 4**, where the highest results were for the proposed model when using the Iraqi model with SG technique and the length of the vector equal to 100.

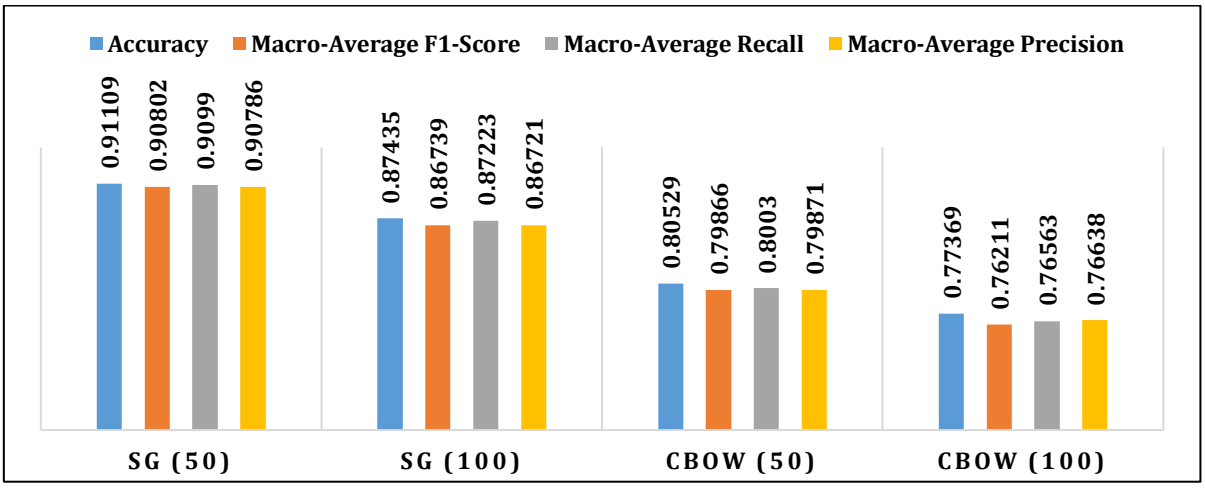

Figure 2. CNN-Two GRU Model Results with IAEDS Dataset.

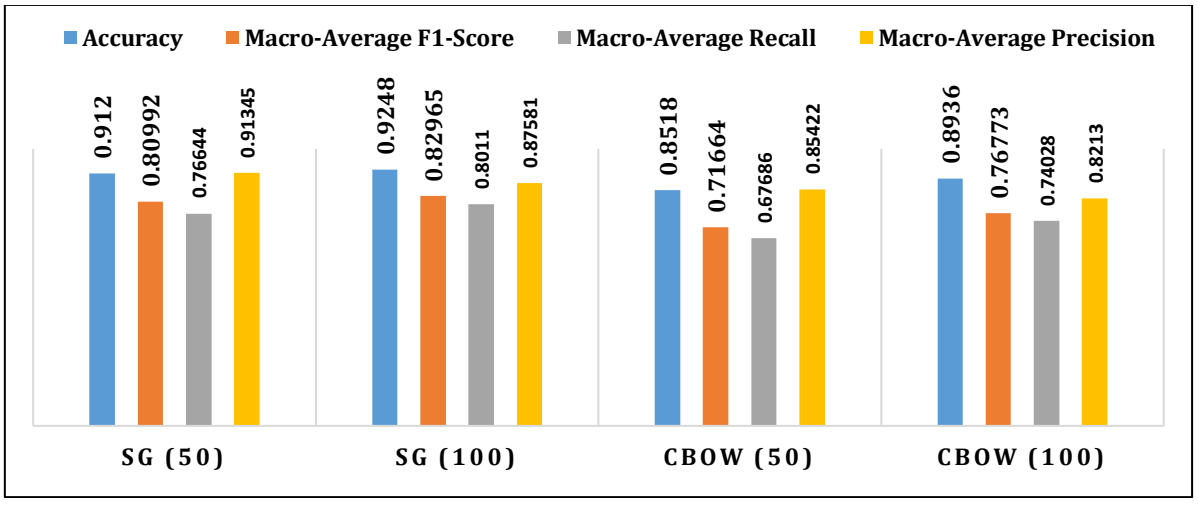

Figure 3. CNN-Two GRU Model Results with ACMID Datasets

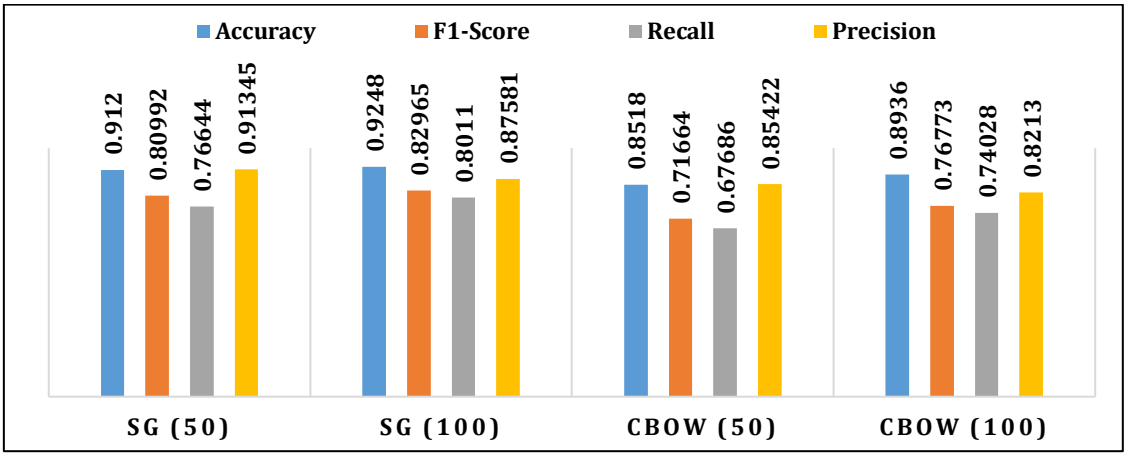

Figure 4. CNN-Two GRU Model Results with IAD Dataset.

True Positive Rate (TPR) and False Positive Rate (FPR) were compared using a Receiver Operating Characteristics (ROC) curve at various threshold settings for the model for each dataset. The ROC curve for the three datasets (IAEDS, ACMID, and IAD) is shown in **Fig. 5**, **Fig. 6**, and **Fig. 7**, respectively.

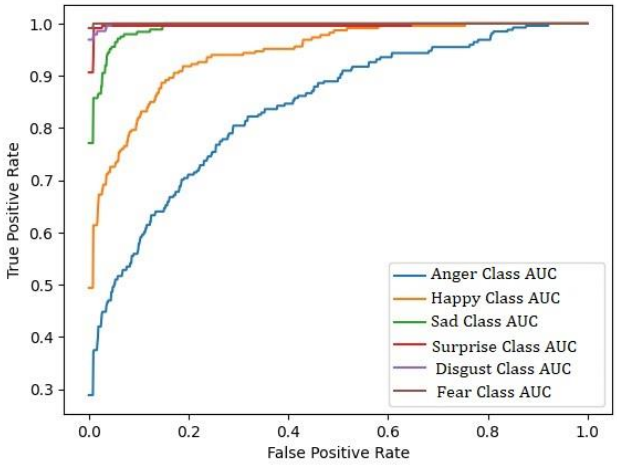

Figure 5. ROC Curve for IAEDS Dataset.

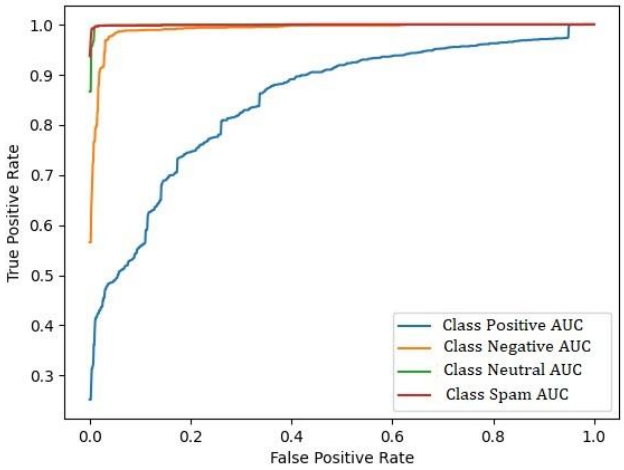

Figure 6. ROC Curve for ACMID Dataset.

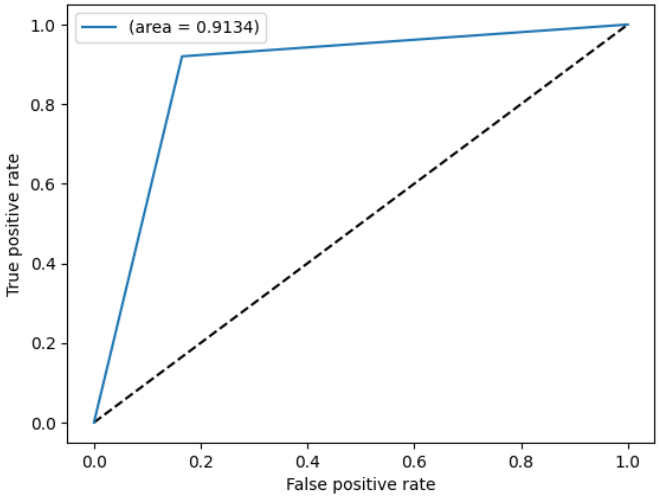

Figure 7. ROC Curve for IAD Dataset.

Results of comparing the model with the previous works concerning the dataset IAEDS presented in **Table 2,** where it outperformed it when using the Iraqi Word2Vec with the SG technology with a length of 50 meters. Comparison results for the IAD dataset are presented in **Table 3**. The model outperformed previous works using the Iraqi model with the SG technique, and the vector length was equal to 100.

Table 2. The proposed CNN-Two GRU Model Comparison with Baseline Models on the IAEDS Dataset.

Model	Accuracy	Macro-averaged		
		F1-score	Recall	Precision
CNN-GRU Model	91.1	0.90802	0.90990	0.90786
Alswaidan N. and Menai M.	87.2	0.64	0.60	0.69
Almahdawi A. and Tehan	87.1	0.61	0.59	0.63

Table 3. Proposed CNN-GRU Model Comparison with Baseline Models on the IAD Dataset.

Model	F1-score	Recall	Precision
CNN-GRU Model	0.92537	0.92079	0.95
ALNAWAS A. and ARICI N.	0.78	0.79	0.82

4. CONCLUSIONS

In this work, the challenge of sentiment and emotion analysis for Iraqi text has been addressed. A hybrid deep learning model is introduced, consisting of one layer of CNN followed by two layers of GRU. Model evaluation experiments were conducted on three Iraqi datasets. For every single dataset, the model performed well. The model also greatly beats its predecessors regarding Accuracy and F1-measure, recall, and accuracy.
Given that the Iraqi datasets were mainly with a small number of samples, we suggest in future works that attention be paid to the Iraqi dialect and that large Iraqi datasets be collected to conduct studies on them.

REFERENCES

Abdullah, M., Hadzikadicy, M., and Shaikhz, S., 2018. SEDAT: sentiment and emotion detection in Arabic text using cnn-lstm deep learning. *Proceedings - 17th IEEE International Conference on Machine Learning and Applications ICMLA,* Orlando, FL, USA, 17-20 Dec, 2018, pp. 835–840. Doi:10.1109/ICMLA.2018.00134

Abdullah, M., and Shaikh, S., 2018. TeamUNCC at semeval-2018 task 1: emotion detection in English and Arabic tweets using deep learning. *NAACL HLT 2018 - International Workshop on Semantic Evaluation, SemEval 2018 - Proceedings of the 12th Workshop*. *New Orleans, Louisiana. Association for Computational Linguistics*, June 2018, pp. 350–357. Doi:10.18653/v1/s18-1053

Abu Kwaik, K., Saad, M., Chatzikyriakidis, S., and Dobnik, S., 2019. LSTM-CNN deep learning model for sentiment analysis of dialectal Arabic. *Communications in Computer and Information Science*, 1108, pp. 108–121. Doi:10.1007/978-3-030-32959-4_8

Alayba, A.M., Palade, V., England, M., and Iqbal, R. 2018a. A combined CNN and LSTM model for Arabic sentiment analysis. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *LNCS* 11015, pp. 179–191. Doi:10.1007/978-3-319-99740-7_12

Alayba, A.M., Palade, V., England, M., and Iqbal, R., 2018b. Improving sentiment analysis in Arabic using word representation. *2nd IEEE International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, London, UK, 12-14 March 2018, pp. 13–18. Doi:10.1109/ASAR.2018.8480191

Almahdawi, A.J., and Teahan, W.J., 2019. A new Arabic dataset for emotion recognition. *Advances in Intelligent Systems and Computing*, 998, pp. 200–216. Doi:10.1007/978-3-030-22868-2_16

Alnawas, A., and Arici, N., 2019. Sentiment analysis of Iraqi Arabic dialect on Facebook based on distributed representations of documents. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *18*(3), pp. 1-17. Doi:10.1145/3278605

Alswaidan, N., and Menai, M.E.B., 2020. Hybrid feature model for emotion recognition in Arabic text. *IEEE Access*, 8, pp. 37843–37854. Doi:10.1109/ACCESS.2020.2975906

Askar, A.K.A.J., and Sjarif, N.N. A., 2021. Annotated corpus of mesopotamian-Iraqi dialect for sentiment analysis in social media. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 12(4), pp. 101–105. Doi:10.14569/IJACSA.2021.0120413

Baali, M., and Ghneim, N., 2019. Emotion analysis of Arabic tweets using deep learning approach. *Journal of Big Data*, 6(1), P. 89. Doi:10.1186/s40537-019-0252-x

Badaro, G., El Jundi, O., Khaddaj, A., Maarouf, A., Kain, R., Hajj, H., and El-Hajj, W., 2018. EMA at semeval-2018 task 1: Emotion mining for Arabic. *NAACL HLT 2018 - International Workshop on Semantic Evaluation, SemEval 2018- Proceedings of the 12th Workshop*, New Orleans, Louisiana. Association for Computational Linguistics, June 2018, pp. 236–244. Doi:10.18653/v1/s18-1036

Ekman, P., 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3–4), pp. 169–200. Doi:10.1080/02699939208411068

Elzayady, H., Badran, K.M., and Salama, G. I., 2020. Arabic opinion mining using combined CNN - LSTM models. *International Journal of Intelligent Systems and Applications*, 12(4), pp. 25–36.

Doi:10.5815/ijisa.2020.04.03

Heikal, M., Torki, M., and El-Makky, N., 2018. Sentiment analysis of Arabic tweets using deep learning. *Procedia Computer Science*, 142, pp. 114–122. Doi:10.1016/j.procs.2018.10.466

Khabour, S.M., Al-Radaideh, Q.A., and Mustafa, D., 2022. A new ontology-based method for Arabic sentiment analysis. Big Data and Cognitive Computing, 6(2), P. 48. Doi:10.3390/bdcc6020048.

Khalil, E.A.H., El Houby, E.M.F., and Mohamed, H.K., 2021. Deep learning for emotion analysis in Arabic tweets. *Journal of Big Data*, 8(1), p. 136. Doi:10.1186/s40537-021-00523-w

Mansy, A., Rady, S., and Gharib, T., 2022. An ensemble deep learning approach for emotion detection in Arabic tweets. *International Journal of Advanced Computer Science and Applications,* 13(4). Doi:10.14569/ijacsa.2022.01304112

Medhat, W., Hassan, A., and Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), pp. 1093–1113. Doi:10.1016/j.asej.2014.04.011

Mohammad, S.M., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S., 2018. Semeval-2018 task 1: affect in tweets. *NAACL HLT 2018 - International Workshop on Semantic Evaluation, SemEval 2018 - Proceedings of the 12th Workshop*, New Orleans, Louisiana. Association for Computational Linguistics, June 2018, pp. 1–17. Doi:10.18653/v1/s18-1001

Mohammed, A., and Kora, R., 2019. Deep learning approaches for Arabic sentiment analysis. *Social Network Analysis and Mining*, *9*(1), pp. 1–12. Doi:10.1007/s13278-019-0596-4

Nassif, A.B., Darya, A.M., and Elnagar, A., 2021. Empirical evaluation of shallow and deep learning classifiers for Arabic sentiment analysis. ACM *Transactions on Asian and Low-Resource Language Information Processing,* 21(1), pp. 1-25. Doi:10.1145/3466171

Ombabi, A.H., Ouarda, W., and Alimi, A.M., 2020. Deep learning CNN–LSTM framework for Arabic sentiment analysis using textual information shared in social networks. *Social Network Analysis and Mining*, 10(1), p. 53. Doi:10.1007/s13278-020-00668-1

Oueslati, O., Cambria, E., HajHmida, M. Ben, and Ounelli, H., 2020. A review of sentiment analysis research in the Arabic language. *Future Generation Computer Systems*, 112(1), pp. 408–430. Doi:10.1016/j.future.2020.05.034

Plutchik, R., 1980. A general psychoevolutionary theory of emotion. In *Theories of Emotion*, pp. 3–33. Doi:10.1016/b978-0-12-558701-3.50007-7

Plutchik, R., 1994. *The psychology and biology of emotion*. HarperCollins College Publishers.

Sailunaz, K., and Alhajj, R., 2019. Emotion and sentiment analysis from Twitter text. *Journal of Computational Science*, 36, p. 101003. Doi:10.1016/j.jocs.2019.05.009

Saleh, H., Mostafa, S., Alharbi, A., El-Sappagh, S. and Alkhalifah, T., 2022. Heterogeneous ensemble deep learning model for enhanced Arabic sentiment analysis. Sensors, 22(10), p. 3707. Doi:10.3390/s22103707

Soliman, A.B., Eissa, K., and El-Beltagy, S.R., 2017. Aravec: a set of Arabic word embedding models for use in Arabic NLP. *Procedia Computer Science*, 117, pp. 256–265. Doi:10.1016/j.procs.2017.10.117

Abdulhameed, T.Z., 2020. Cross language information transfer between modern standard Arabic and its dialects – a framework for automatic speech. Western Michigan University. https://scholarworks.wmich.edu/dissertations

Wint, Z.Z., Manabe, Y., and Aritsugi, M., 2018. Deep learning based sentiment classification in social network services datasets. *Proceedings - 2018 IEEE/ACIS 3rd International Conference on Big Data, Cloud Computing, Data Science and Engineering, BCD 2018*, Yonago, Japan, 12-13 July 2018, pp. 91–96. Doi:10.1109/BCD2018.2018.00022

Word2vec Embeddings. https://radimrehurek.com/gensim/models/word2vec.html