Multimodal Sentiment Analysis of Social Media Posts Using Deep Neural Networks

Aria Naseri Karimvand, Shahla Nemati*, Reza Salehi Chegeni, Mohammad Ehsan Basiri
Department of Computer Engineering, Shahrekord University, Shahrekord, Iran.
aria.nk@yahoo.com, s.nemati@sku.ac.ir, reza.sch@yahoo.com, basiri@sku.ac.ir

Received: 2021/06/03 Revised: 2021/09/07 Accepted: 2021/09/16

***Abstract*— With the fast growth of social media, they have become the most important platform for posting multimodal content generated by users. Much of the data on social networks such as Instagram and Telegram is multimodal data. With the aim of analyzing such multimodal data in social networks, multimodal sentiment analysis has become one of the most significant subjects for researchers in the field of emotion recognition and data mining. Although multimodal sentiment analysis of social media data for English language has been addressed in several researches recently, few studies addressed the problem for the Persian language which is the official language of more than 120 million of people around the word. In this study, a multimodal deep learning model is proposed to address this problem. The proposed method utilizes a bi-directional long short-term memory (bi-LSTM) for processing text posts and a VGG16 convolutional network for analyzing images. A new dataset of Instagram and Telegram posts, MPerSocial, containing 1000 pairs of images and Persian comments is introduced in the current study and used for evaluating the proposed method. The results of experiments show that using the fusion of textual and image modalities improves sentiment polarity detection accuracy by 20% and 8% compared with the scenario in which image and text modalities in isolation. Also, the performance of the proposed model is better than three similar deep and four traditional machine learning models. All codes and dataset used in the current study are publicly available at GitHub.**

Keywords*— *Social Networks; Persian Language; Sentiment Analysis; Deep Learning; Instagram Posts.

1. INTRODUCTION

Nowadays, social media platforms including Instagram, twitter, and Telegram have turned into valuable sources of information for both people and organizations [1]. Using these platforms, users can contact with their physically distant friends and share information. These platforms have billions of users and produce a huge amount of data every minute. People who post or comment using these platforms have different culture, age, gender, attitude, and emotion. This makes the processing of data generated in such platforms a challenging task. Processing such data may be performed with different goals including analyzing users experience [1], advertising [2], education [3], predicting personality traits of users [4], and sentiment analysis (SA) [5].

Knowing the sentiment of online users can be useful for people working in various fields such as economics, marketing, and politics [5], [6]. Sentiment analysis is a subset of data mining (DM) and natural language processing (NLP) that attempts to extract or categorize sentiment of social media automatically [5]. These sentiments can be either positive or negative [7]. Sentiment analysis (SA) has many applications in extracting customers' opinions about a particular product or service [7], extracting emotions from political contexts [6], understanding people's attitude toward environmental issues [8], and recognizing the flow of emotions in social media [5]. Sentiment analysis (SA) can be performed at three levels [5], [7]. Document level which extracts the sentiment of a text as a whole [7], sentence level which considers sentiment sentence by sentence [7], and aspect level which focus on different aspects of a product or service [9].

In the past, sentiment analysis was mostly based on text-based data, but today, with the proliferation of social networks, multimodal sentiment analysis has become a very important issue [10], [11]. With the rapid development of social networks, these networks have become the most important platform for sending multimodal content generated by users [11]. A large part of the data in social networks such as Instagram and Telegram is multimodal data (e.g., image and text). Traditional sentiment analysis methods were usually performed with the aim of analyzing sentiment on textual data [5], but due to the increasing growth of multidimensional data, multimodal sentiment analysis has become one of the most important issues for researchers in the field of data mining [10].

The accuracy of multimodal systems relies on the data fusion mechanism used to combine the information carried out by different modalities [13], [14]. Several fusion methods have been utilized in the previous studies, including canonical correlation analysis [14], latent space [13], hierarchical fusion [16], and tensor fusion [26]. These methods differ in the level at which the fusion takes place and in the internal mechanism of the fusion method [13]. For example, the fusion may be performed at the score-level or feature-level; in the former the scores produced by the classification methods are combined, while in the latter the features extracted from different modalities are first combined then fed into the classifier to produce the final classification results [13].

Today, many researchers, especially for the English language, conduct several researches in the field of multimodal data analysis [13]. Nevertheless, despite the expansion of Persian language users in social networks and the production of millions of posts by them, few researches on multimodal sentiment analysis study have been reported in the Persian language. In multimodal SA, complementary modalities such as image or video modalities are exploited beside the text modality [13]. Multimodal SA is more challenging than traditional text-based SA because it needs appropriate modeling of the relation between the modalities [12]. This becomes more challenging when modalities are not

10.22133/ijwr.2021.289091.1096 20.1001.1.26454335.2021.4.1.1.7

synchronized [13]. In such cases, more complicated fusion mechanisms are needed [14] deep learning usually offers promising approaches to such conditions [11], [12].

Most studies for multimodal SA targeted the English language [15], [16] and there is only one study [17] that directly addressed Persian multimodal SA. Persian is an Indo-European language spoken by more than 120 million people in Iran, Tajikistan and Afghanistan [18], [19]. As pointed out in several studies Persian language is a challenging language for NLP applications [19]-[21]. Some difference between the Persian and English languages are related to the structure of plurals, negative, and modal verb formation [21]. Moreover, there are limited linguistic resources such as datasets, standard lexica, and word embeddings for Persian, making this language a low-resource language for NLP applications [21], [21].

Existing studies for sentiment analysis of Persian language can be categorized into three main categories: lexicon-based [24], [25], ML-based method [26], [27], and deep learning-based methods [28], [30]. The first category, utilized a predefined dictionary of sentiment words and their corresponding sentiment polarity or score to compute the sentiment of a chunk of text [24]. Machine learning methods, on the other hand, need a labeled training dataset to train a classifier for labeling unseen texts [26]. Deep models which can be considered as a subset of ML methods, usually need large training data and achieve higher accuracy compared to previous two methods. However, analyzing Persian Instagram and Telegram posts only pointed out in [30] where a new dataset of Instagram post, Insta-text were introduced. This dataset contains 9000 Instagram posts and their corresponding sentiment polarity label (i.e., positive, negative, and neutral). This study only applied Word2Vec and did not report the results of sentiment classification [30].

A new deep learning model is proposed in the current study for sentiment analysis of Persian multimodal Instagram and Telegram posts to address the above-mentioned problem. The proposed model has an image branch for extracting visual features from images in the posts, and a text branch for extracting textual features from the posts. For the image branch, the VGG16 network which consists of 16 convolutional layers is used and for the text branch, a bi-directional long short-term memory (bi-LSTM) is used. The main advantage of using the VGG16 is that it can extract different features by replacing larger convolution kernels with small consecutive 3×3 convolution kernels [30]. On the other hand, bi-LSTM has the ability of extracting sequential dependencies in both forward and backward direction. To make a fusion of the high-level text and image features, the outputs of text and image branches are concatenated and passed to a fully connected layer for making the final classification.

The current study has the following contribution in multimodal SA domain:

1) A new Persian social media dataset (MPerSocial) with sentiment polarities suitable for multimodal sentiment analysis is proposed.

2) A new deep learning model for analyzing sentiment of Persian social media posts is proposed.

3) The first multimodal deep fusion model for sentiment analysis of social media posts is presented.

The paper continues as follows: Section 2 briefly reviews related studies in Persian SA and multimodal SA. Section 3 describes the details of our MPerSocial dataset. In Section 4, more details about the structure of the proposed method are presented. Experimental results and a brief discussion of them is presented in Sections 5. Finally, conclusions and some future work directions are presented in Section 6.

2. LITERATURE REVIEW

This section is divided into two subsections; in the first subsection, we review the most relevant recent studies in Persian SA and in the next subsection, we present a brief review of recent related multimodal SA studies.

2-1. Persian sentiment analysis

In [26], two datasets BG-Data and BC-Data have been used which contain the opinions of Persian language users about mobile phones. The BG-Data dataset contains short user comments about mobile phones, and the BC-Data dataset, extracted from a popular Persian-language site for digital products, seeks to address the problem of short comments in the BG-Data dataset. They then provided a lexicon-based framework to address challenges in the Persian language, such as misspellings, stemming, and the use of informal language by users. Finally, to evaluate the effectiveness of their proposed framework, machine learning algorithms were used to analyze sentiment in this data. In [7] a lexicon-based sentiment analysis method was used and a Persian lexicon was presented. The lexicon consists of 1500 words besides their polarity and sentiment intensity. The words that make up the dictionary were chosen from adjectives, adverbs, verbs, and nouns. In this research, to evaluate the proposed lexicon, support vector machines (SVM) and naïve Bayes algorithms were used to classify texts. SVM algorithm with 69.54% compared to naïve Bayes algorithm with 65.02% had better performance in classification, based on all features.

In [31], for the first time, a data set containing Persian-speaking users' opinions about Iranian films was introduced. Then, multilayer perceptron-based models, automatic encoders, and convolutional networks were used to classify the data. Auto-encoder multilayer network performed better than multilayer perceptron, but convolutional neural network (CNN) model performed best at 82.86%. In [32], a hybrid model, including convolutional network and LSTM was used to classify the sentiment of textual data in two Persian data sets. The data sets used were collected from Digikala website and Twitter. The reason for using the hybrid method was that the convolutional network is suitable for extracting local features from the text and the recurrent network is suitable for extracting sequential dependencies from the text. Finally, the convolutional and recurrent hybrid neural network was compared with other models and it was found that this hybrid network has a better performance than other models in classifying sentiment in two data sets.

In [34], the authors have analyzed emotions using a data set related to an electronic services website. The proposed method was based on deep convolutional models and LSTM. Also, active learning was used to increase the accuracy of the model. The researchers in [35] introduced a Persian data set for the analysis of sentiment and then, due to the small number of data, they increased the data with augmentation methods. The data set was labeled in two modes; binary and three-class.

Finally, traditional machine learning and deep learning algorithms were used to classify the data. They showed that in binary mode, the best performance among different algorithms was achieved by the SVM with 91.31% and bi-LSTM using embedding layer with 91.98%. In the three-class mode, among the algorithms, the SVM with 67.62% and bi-LSTM using FastText embedding with 69.33% were the best algorithms.

In [38], a data set containing Persian tweets and their transliteration has been introduced. This data set has three classes. In order to use non-Persian data, they first validate the word with the help Persian words related to Wikipedia, then translated the word using a translator. Then, they embed words using the Bert embedding and classify them using a model with three bi-LSTM models. Finally, the proposed algorithm was compared with two machine learning algorithms, namely naïve Bayes and random forest, which according to the results, the proposed algorithm has been able to achieve better accuracy. In [40], a data set related to Persian users' opinions about hotels, with two labels, positive and negative, has been introduced. After collecting the data set, they normalized the data using tokenization, normalization, and stemming, then used Word2vec to embed the words. In this research, a hybrid classification method is used to classify the data. This hybrid method includes machine learning algorithms and deep learning models. Finally, they concluded that the proposed hybrid classification method has performed better than traditional machine learning and deep learning methods.

2-2. Multimodal sentiment analysis

In [38], the authors introduced two data sets containing user tweets and analyzed multimodal sentiment in this data. The first data set was labeled by one person, but the second data set was labeled by three people. In this study, a convolutional network was used for text and image data, and a method called Multi-NN was used to merge the two networks. The Multi-NN method had two approaches of early and late optimizations. In the early approach, the outputs of the text and image layers were first merged separately with a fully connected layer, then the output of the fully connected text and image layers were merged. In the late approach, they first used two fully connected layer for text and two similar layers for image data. Then, they merged the output of these four fully connected layers together. The reported results showed that the late method outperformed the early method in binary mode.

In [39], a multimodal data set related to users' opinions about different restaurants was introduced. In this research, a model containing Vgg16 network for image and a Bi-GRU network for text was introduced. According to the obtained results and the performance of the proposed model compared to other models, it was shown that the features extracted from the images after combining with textual features have been able to increase the performance of the proposed model. In [40], a data set belonging to the social network Instagram was collected which includes a pair of images and text. Then, they used a ResNet network for the image and a recurrent network for the text to analyze the sentiment in this multimodal data, and finally combined the text and image networks and concluded that the combination of these two networks improves the performance of the model.

In [41], a multifunctional approach for analyzing multimodal sentiment in two data sets has been introduced. Given that the pair of images and texts may not always be present in the data, the proposed approach tried to solve this problem. This approach consisted of three classifiers, one for text analysis, one for image analysis, and one for prediction, based on a combination of both methods. Single-state classifications help the model to make accurate predictions as long as there is no image or text pair. In [42], using a data set containing pairs of images and text and typography data, multimodal sentiment analysis was performed. The data set was related to user tweets in posts related to Indian criminal courts. In the proposed method, first the input tweets were checked in terms of type, then an appropriate pre-processing was performed to separate the text from the image in typography by OCR method. Finally, using deep learning, they analyzed the sentiment in this data and the proposed method showed good performance with 91.32% accuracy.

In [43], the authors introduced a collection of data related to users' tweets about natural disasters with the category of information and humanitarian. In this study, to analyze the sentiment in this data, they used VGG16 for image mode and convolutional network for text mode. Then, they combined the output of these two networks and produced the output by a classifier. Finally, the accuracy of the model in information data category reached 84.4%, while in the humanitarian data it reached to 78.4%.

An overview of the above-mentioned studies is shown in Table I. In summary, although deep learning has been applied to Persian sentiment analysis, previous studies in Persian sentiment analysis have not addressed multimodal analysis. Moreover, no publicly available dataset exists for multimodal sentiment analysis on the Persian language. The current study addresses these problems by introducing MPerSocial dataset and proposing a multimodal deep learning model for Persian sentiment analysis.

3. MPERINST DATASET

3-1. Collecting the dataset

In this research, for the first time, a multimodal data set in Persian language is collected and introduced. This data set includes public posts (pairs of images and text) of Persian language users on Instagram and Telegram social networks, and we name it MPerSocial[1]. Fig. 1 shows two examples of MPerSocial image and text pairs. MPerSocial consists of 1000 posts on social networks with emotional subjects, whose data is labeled as negative or positive polarity. From 1000 posts, 439 were labeled as negative and 561 were labeled as positive.

Part of the MPerSocial data set is collected using the social networking API and another part is collected manually. In the introduced data set, the images have different dimensions and formats, which needs some preprocessing steps addressed in the next section. Also, the texts have different lengths in terms of words, which are shown in Fig. 2. In order to use the introduced data set, it is necessary to solve the problems expressed in images and texts, such as different dimensions or different formats in images, so that they can be used as input to the neural network. As shown in Fig. 2, the length of most of the texts in the dataset is in the ranges of up to 10 and 10 to 20 words, and the average length of the sentences in the whole dataset is 14.21.

[1] https://github.com/mebasiri/Multimodal-Persian-SA

TABLE I. AN OVERVIEW OF RELATED STUDIES FOR PERSIAN AND MULTIMODAL SA.

Domain / Study	Persian SA	Multi-modal SA	Description
Basiri et al. [25]	✓		Two datasets were introduced and a lexicon-based framework
Basiri et al. [19]	✓		A lexicon-based method was proposed
Dashtipour et al. [32]	✓		A dataset was introduced and some deep models were tested.
Bokaee Nezhad et al. [33]	✓		A hybrid CNN LSTM model was proposed.
Ashrafi Asli et al. [34]	✓		Deep models aand ctive learning was used.
Sharami et al. [35]	✓		Traditional ML and deep moels wee evaluated.
Sabri et al. [36]	✓		A data set of Persian tweets and their transliteration was introduced
Dashtipour et al. [37]	✓		A data set related to Persian users' opinions was introduced.
Xu et al. [38]		✓	Two data sets containing user tweets were introduced.
Truong et al. [39]		✓	A model containing Vgg16 network for image and a Bi-GRU network for text was introduced.
Kruk et al. [40]		✓	A data set was collected from Instagram. A ResNet network for the image and a recurrent network for the text was proposed.
Fortin et al. [41]		✓	A multifunctional approach for analyzing multimodal sentiment in two data sets has been introduced
Kumar et al. [422]		✓	Tweets in posts related to Indian criminal courts were analyzed
Ofli et al. [43]		✓	VGG16 for image mode and convolutional network for text modewere use.

3-2. *Preprocessing*

Due to the fact that the data were collected from sources that did not consider the data mining process and did not have a proper structure, the data needed to be converted into suitable data for injecting into the neural network using appropriate preprocessing methods. In this research, considering that each sample of the presented data set contains a pair of images and text, it is necessary to perform preprocessing operations on images and texts separately.

The images in the data set have different dimensions and formats. It is necessary for the data that should be imported to the network to have the dimensions and format. Initially, the dimensions of the images were changed to 244 by 244 per pixel unit, then due to the fact that the images had "PNG" and "JPG" formats, to make the format of all images similar, we convert them to "JPG". Finally, the input tensor for the images is made. this is a 4-dimensional tensor and includes the number of images, width, height, and number of channels. Due to the fact that the images are in RGB format, the input images have 3 channels. As the collected text data does not have a suitable

Image		
Persian text	در قلب خود باور داشته باشید که یک حادثه شگفت انگیز قرار است اتفاق بیفتد باور کنید همان می‌شود که باور دارید عاشق زندگی‌تان باشید.	در زندگی، وقتی کاری برای دوست داشتن یا انگیزه ای برای امید داشتید، بدانید که فرد شادی خواهید بود.
English translation	Believe in your heart that an amazing event is about to happen. Believe that what you believe will happen, love your life.	In life, when you had something to do to love or a motivation to hope, know that you will be a happy person

Fig. 1. Two samples of image-text pair in MPerSocial dataset.

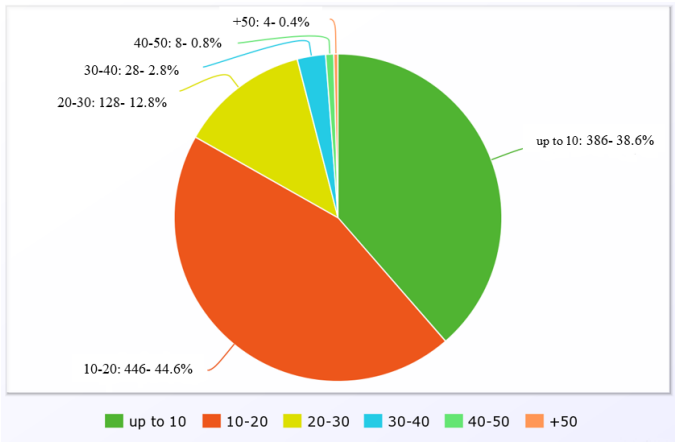

Fig. 2. The distribution of post length in MPerSocial data set.

structure to use in the neural network, pre-processing operations need to be performed on them. Unfortunately, there is no suitable and accurate library in Persian for pre-processing Persian texts, and existing libraries, such as Hazm, also have problems such as the inability to delete stop words. Accordingly, in order to remove the stop words and delete additional signs and phrases, we separately compiled a list including Persian stop words and un-necessary words and phrases. In the next step, by applying them to the data set, we normalized the data.

4. PROPOSED MODEL

The architecture of the proposed deep model is shown in Fig. 3. As shown in the figure, the input to the system is an image and its corresponding text written by the same user. Following this layer, two separate deep neural networks are used to extract visual and image features from the inputs. The outputs of these layers are then sent to a concatenation module for making a feature vector for the final classification. The internal structure of the image and text deep networks is shown in Fig.4. More details about the deep neural networks and other modules will be presented in the following sub-sections.

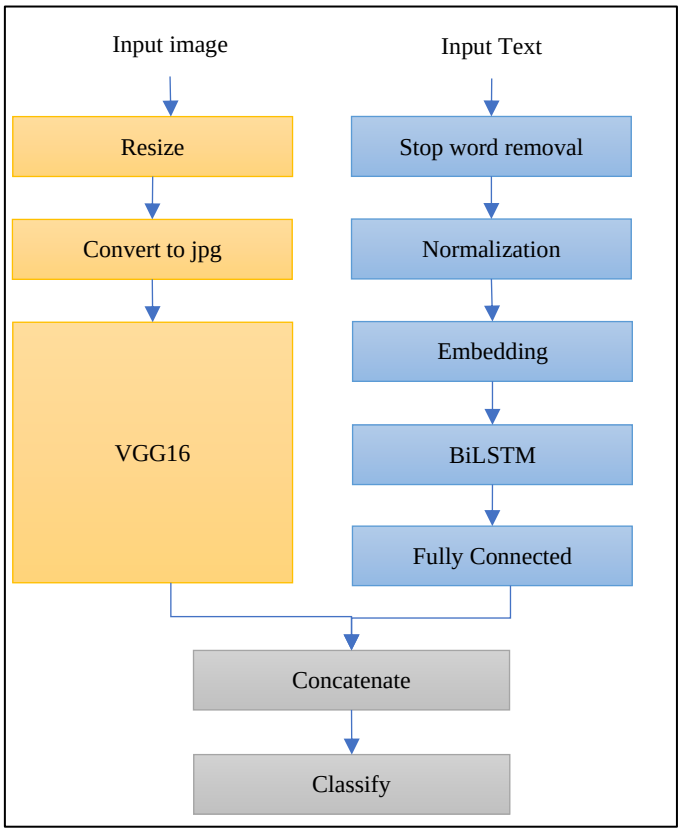

Fig. 3. Overall structure of a the proposed model.

4-1. Text processing

A word embedding layer is employed to convert the text into a numerical vector. The input's length and the embedding dimension are 20 and 10, respectively. In order to model the sequential relation between the words, following the embedding layer, a bidirectional LSTM layer containing 10 cells is used. The bi-LSTM layer considers both forward and backward relation in the text using LSTM cells which internally are as follows:(Eq. (1) to (5))

$$f_t = \sigma_g(w_f\, x_t + U_f h_{t-1} + b_f) \quad (1)$$

$$i_t = \sigma_g(w_i\, x_t + U_i h_{t-1} + b_i) \quad (2)$$

$$o_t = \sigma_g(w_o\, x_t + U_o h_{t-1} + b_o) \quad (3)$$

$$c_t = f_t\, o\, c_{t-1} + i_t\, o\, \sigma_c(w_c\, x_t + U_c h_{t-1} + b_c) \quad (4)$$

$$h_t = o_t\, o\, \sigma_h(c_t) \quad (5)$$

where, f_t, i_t, o_t, c_t, and h_t are forget, gate input, gate output, condition, and output vectors. σ_g and σ_c are sigmoid and Tanh activation functions.

The Bi-LSTM layer consists of two independent LSTMs that summarize information in both directions. Finally, the output of the Bi-LSTM layer enters a fully connected layer, which includes 1000 neuron and a batch normalization layer. The fully connected layer neuron uses a ReLU activator. In the fully connected layer of the text branch, 1000 neurons are used because in the image branch, the output of the final neuron is 1000, and in the time of merging, it is required for the two branches to have same size.

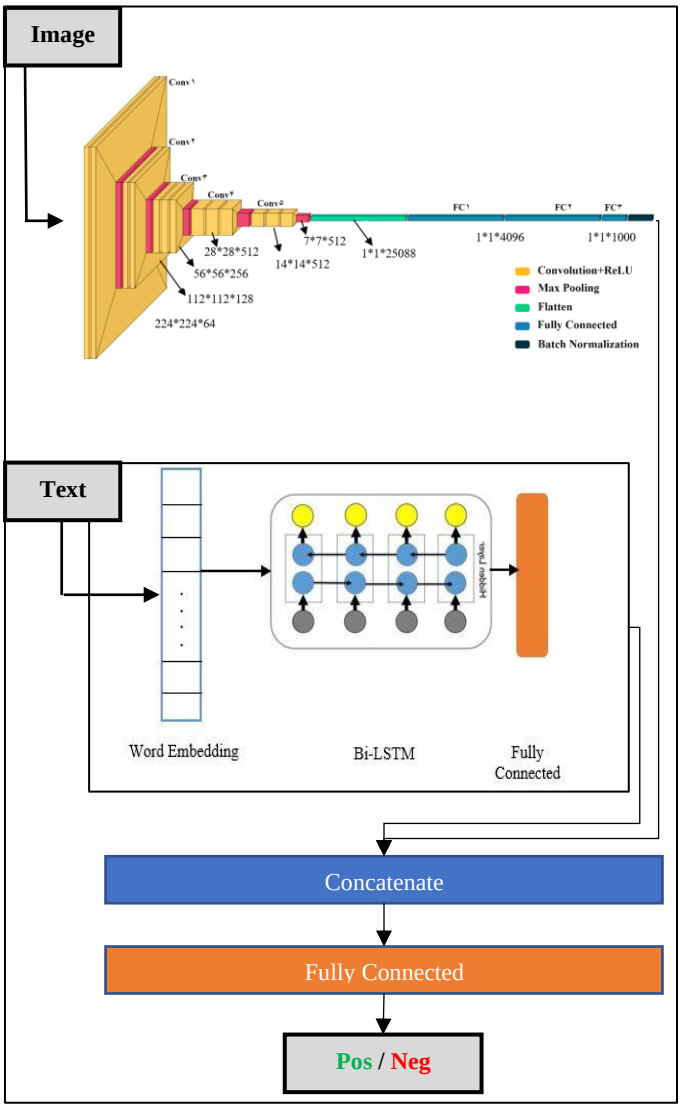

Fig. 4. Overall structure of a the proposed model.

4-2. Image Processing

Due to the high performance of VGG16 architecture in image classification, in this research, this architecture is used to teach the network's image branch. VGG16, one of the most advanced pre-trained deep learning architectures, was first introduced in the ImageNet competition. This architecture has a number of convolutional layers, behind which there are pooling layers that make the layers shrink. Given that this architecture is already trained on the ImageNet data set, many researchers are re-training it to make changes to their databases to take advantage of its capabilities.

The input to the VGG16 network is a 4-dimensional tensor and is first passed from the Conv1 block, which consists of 2 convolutional layers with 64 filters and a 3×3 kernel size, plus a 2×2 maximum pooling layer with stride size 2. This pooling layer is used for sampling and reducing the dimension of features. Then, the output of the first block enters the second block, Conv2, this block also includes two convolutional layers with 128 filters and one layer of maximum pooling with the specifications similar to that described for the previous block. The output of the second block also enters the Conv3 block, which consists of three convolutional layers with a 256 filter,

which is reduced by one layer of maximum pooling of its feature dimensions. Then, there are two Conv4 blocks and Conv5, each of which consists of 3 convolutional layers with 512 filters, a 3×3 kernel, and a maximum pooling layer.

It is observed that after passing the image of each block, its dimensions decrease and its depth increases. All convolutional blocks use the ReLU activator, which maps outputs smaller than zero to zero and larger than zero to themselves. Then, after the convolutional blocks, their output is sent to a flatten layer to convert the output of neurons into a 1D vector. Then, the output of this layer is connected to the first and second fully connected layers, respectively. They contain two layers with dimensions 4096 and ReLU activators. In the third fully connected layer, there is a 1000-dimensional layer with ReLU activator. This is due to the fact that the model uses the weights of the ImageNet data set which has 1000 classes. In order to teach this model on the data set introduced in this study, changes are made in the output neurons after integration. Finally, a batch normalization layer is used to increase the training speed and model accuracy.

Finally, the outputs of both the image and text branches are entered as input in the concatenate layer to be combined with each other to form a common vector. The output of the concatenate layer, which is a 2000-neuron layer, is then sent to the fully connected layer to produce the final output. The fully connected layer includes a batch normalization and two dense layers with three dropouts between them. The reason for using dropout is to prevent the model from overfitting in training time. Also, since the model ultimately needs to produce a negative or positive output for each input, in the last part of the fully connected layer, a dense layer is used to produce the result using a sigmoid function, which produces its output in the range 0 or 1.

5. Experiments and Results

In order to train the proposed system, Adam optimizer with binary cross-entropy loss function and accuracy metric were used. Moreover, 20 epochs were used with batch size of 32. In order to carry out the experiments we used a machine with two Intel(R) Xeon(R) 2.00GHz CPUs with 6MB cache, 13GB RAM, and a Tesla K80 GPU with 12GB GDDR5 VRAM. All codes were written in Python 3.6 in Google Colab environment using Keras library [44].

5-1. Comparison with Deep models

In Fig. 5, the proposed model is compared with the following similar deep models as follows:

1) VGG16-GRU

This model is very similar to the proposed model except for the text branch which here, bi-LSTM is replaced by GRU. In the GRU model, as in the previous model, we first enter the length of the sentences and then use a word embedding layer to create a numerical dense vector, then there is a GRU layer with 10 cells that its output is connected to a dense layer with 1000 neurons in the output. The output of this layer is connected to a batch normalization layer that is responsible for speeding up the network. Finally, using a one neuron layer, the final is produced.

2) 2CNN-bi-LSTM

This model is similar to the proposed model except for the image branch which is replaced by a two-dimensional convolutional model. This model consists of two convolutional layers followed by a maximum pooling layer at first, then a dropout layer with a rate of 0.5, which is responsible for preventing overfitting in the network. After these layers, there are flatten, dense, and batch normalization layers. Finally, using a dense layer with one neuron at the output and a Sigmoid activator, the results of sentiment classification in images are produced.

3) 2CNN-GRU

This model is the combination of the previous two models in the following way. The image branch is a two-dimensional CNN and the text branch is a GRU layer as described above.

The results of comparing the accuracy of these models with the proposed model in the training and test time are shown in Fig. 5. As showed in the figure, the proposed model outperforms the three above-mentioned deep models in terms of accuracy in the test time. Moreover, the performance of the proposed model (Fig. 5 a) and the VGG16-GRU model (Fig. 5 b) are better than the other two models. This may be the effect of using VGG16 model in the proposed model and VGG16-GRU model for classifying the images. Comparison of the loss of the proposed method with the three similar deep models is shown in Fig. 6.

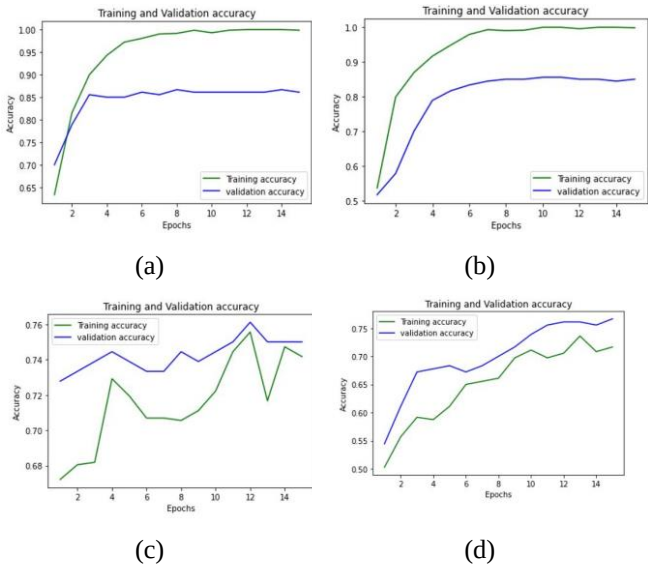

Fig. 5. Comparison of the accuracy obtained using the proposed method (a) with VGG16-GRU (b), 2CNN-bi-LSTM (c) and 2CNN-GRU (d) methods on the MPerSocial dataset.

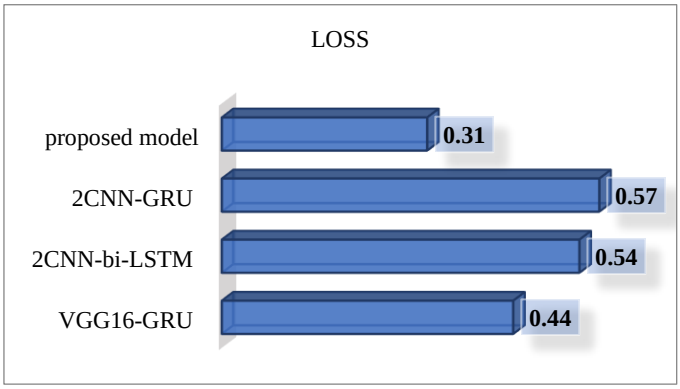

Fig. 6. Comparison of the loss for the proposed method with VGG16-GRU, 2CNN-bi-LSTM, and 2CNN-GRU methods on the MPerSocial dataset.

As showed in Fig. 6, the loss of the proposed method is lower than the other three methods. This makes the proposed method more suitable for multimodal SA of posts in comparison with similar models described above.

5-2. Comparison with ML models

Four classical machine learning algorithms are used to classify sentiment in images, text, and multimodal mode for comparison, namely SVM algorithm, random forest, decision tree, and naïve Bayes. In the SVM algorithm we tested RBF, polynomial, and linear kernels and reported the best results obtained which belongs to RBF. For the random forest algorithm, we used Gini impurity to measure the quality of a split, best policy in the splitter, and 2 as the minimum number of samples required to split an internal node. For the random forest, we used 100 as the number of trees in the forest, Gini as the measure of the quality of a split, and 2 as the minimum number of samples required to split an internal node. For the naïve Bayes, we used the GaussianNB method with default parameters in Keras library.

In order to show the effect of combining images and texts in machine learning algorithms, we use machine learning algorithms to classify sentiment separately in images and texts and in the combination of these two modes and compare the obtained accuracy with that of deep models in Table II.

As showed in Table II, the best performing ML method is random forest with 74% accuracy which is 17% lower than the accuracy of the proposed model. This shows the higher ability of the proposed model in classifying sentiment in multimodal data. Moreover, in almost all cases the accuracy of using both image and text is higher than the accuracy of utilizing each modality in isolation. This emphasizes the benefit of using multimodal data for identifying sentiment of users in social media.

5-3. Discussion

In order to further analyze the effect of using text and image modalities in multimodal sentiment analysis, we measured the percent of obtaining the same results by image and text modalities using the proposed method. The results showed that for 41% of test samples, both modalities predict the same sentiment polarity, while for 59% of test samples their prediction were not the same. This, along with the results reported in Table II. Shows that the combination of text and image modalities improves the accuracy of the model. Two samples for which the proposed model produced different sentiment polarity using the text and image modalities are shown in Fig. 7.

6. CONCLUSION

The increasing growth of the Internet and online activities such as chats, transactions, and e-commerce has led many researchers in the field of sentiment analysis to move to the analysis of sentiment in these areas. On the other hand, with the recent development of social networks and due to their high capabilities, many users who want to share their opinions on various topics, publish posts that usually have an image and a text that complements the user's opinion about the image. sentiment analysis in these multimodal user posts can have many applications in various fields, including medical care to diagnose stress, anxiety, and depression. Many studies in the field of multimodal sentiment analysis using deep neural networks have been conducted in English. In this study, for the first time, using a deep neural network, we analyzed sentiment on the data of Persian-speaking users of social networks. In this study, we have used four deep neural network models, namely VGG16-bi-LSTM, VGG16-GRU, 2CNN-bi-LSTM, and 2CNN-GRU. In order to compare the use of different dimensions and the effect of the multimodal model, we obtained the results of different dimensions separately and compared with the multimodal model. Then, after this step, to show the superiority of the proposed deep neural network over traditional machine learning models, we have compared the proposed model with four traditional machine learning models. The results showed that the proposed deep multimodal model outperforms traditional ML models significantly.

TABLE II. COMPARISON OF THE ACCURACY OBTAINED USING THE PROPOSED MODEL, FOUR ML ALGORITHMS, AND THREE SIMILAR DEEP MODELS.

Modality / *Method*	***Text***	***Image***	***Multimodal***
Naïve Bayes	53%	61%	61%
SVM	56%	62%	66%
Random Forest	70%	70%	74%
Decision Tree	58%	62%	64%
VGG16-GRU	69%	83%	85%
2CNN-bi-LSTM	71%	71%	76%
2CNN-GRU	69%	71%	74%
Proposed model	71%	83%	**91%**

Image		
Predicted Polarity by image modality	Negative	Positive
Persian text	زیبا؛ هوای حوصله ابری ست !!!...	کسی که بدونه شما ناراحتید و بذاره با اون حال روزهاتون رو سپری کنید هیچ کستون حساب نمیشه...
English translation	beautiful; The weather of the mood is cloudy... !!!	Someone who knows you are upset and lets you spend your time in that way has nothing to do with you ...
Predicted Polarity by text modality	Positive	Negative

Fig. 7. Two image-text pairs for which the image and text networks predict different polarity.

Also, for the first time, a collection of data related to the opinions of Persian-speaking users on two social networks, Instagram and Telegram was introduced. Each instance of this dataset contains a pair of related images and text. The data in this dataset has positive or negative polarity labels. Due to the limitations of using APIs, we collected the data manually as well as using the API. On the other hand, due to the nature of the data, the provided labels are limited to the sentiment polarity of the data, which could be improved to more fine-grained sentiment in the future. One of the most important needs of deep neural networks is large data sets. In neural networks, the larger the data set, the better the performance of the network. Hence, developing such large multimodal data sets may be considered as a future work. Also, to improve the quality of sentiment analysis in the text, one can use pre-training embeddings such as word2vec and FastText. Furthermore, there is other data such as video and audio. Sometimes, users share posts that contain a pair of video and text or a pair of audio and text, which can be considered for future work. This adds more dimensions for analyzing sentiment.

REFERENCES

[1] Rhee, Bo-A., Federico Pianzola, and Gang-Ta Choi. "Analyzing the museum experience through the lens of Instagram posts." Curator: The Museum Journal (2021).

[2] Ershov, Daniel, and Matthew Mitchell. "The Effects of Influencer Advertising Disclosure Regulations: Evidence From Instagram." In Proceedings of the 21st ACM Conference on Economics and Computation, pp. 73-74. 2020.

[3] Romero-Rodríguez, José-María, Carmen Rodríguez-Jiménez, Magdalena Ramos Navas-Parejo, José-Antonio Marín-Marín, and Gerardo Gómez-García. "Use of Instagram by Pre-Service Teacher Education: Smartphone Habits and Dependency Factors." International Journal of Environmental Research and Public Health 17, no. 11 (2020): 4097.

[4] Ferwerda, Bruce, Markus Schedl, and Marko Tkalcic. "Predicting personality traits with instagram pictures." In Proceedings of the 3rd Workshop on Emotions and Personality in Personalized Systems 2015, pp. 7-10. 2015.

[5] Xing, Frank, Lorenzo Malandri, Yue Zhang, and Erik Cambria. "Financial Sentiment Analysis: An Investigation into Common Mistakes and Silver Bullets." In Proceedings of the 28th International Conference on Computational Linguistics, pp. 978-987. 2020.

[6] Khatua, Aparup, Apalak Khatua, and Erik Cambria. "Predicting political sentiments of voters from Twitter in multi-party contexts." Applied Soft Computing 97 (2020): 106743.

[7] Basiri, Mohammad Ehsan, and Arman Kabiri. "HOMPer: A new hybrid system for opinion mining in the Persian language." Journal of Information Science 46, no. 1 (2020): 101-117.

[8] Abdar, Moloud, Mohammad Ehsan Basiri, Junjun Yin, Mahmoud Habibnezhad, Guangqing Chi, Shahla Nemati, and Somayeh Asadi. "Energy choices in Alaska: Mining people's perception and attitudes from geotagged tweets." Renewable and Sustainable Energy Reviews 124 (2020): 109781.

[9] Ma, Yukun, Haiyun Peng, Tahir Khan, Erik Cambria, and Amir Hussain. "Sentic LSTM: a hybrid network for targeted aspect-based sentiment analysis." Cognitive Computation 10, no. 4 (2018): 639-650.

[10] Tran, Ha-Nguyen, and Erik Cambria. "Ensemble application of ELM and GPU for real-time multimodal sentiment analysis." Memetic Computing 10, no. 1 (2018): 3-13.

[11] Poria, Soujanya, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. "Fusing audio, visual and textual clues for sentiment analysis from multimodal content." Neurocomputing 174 (2016): 50-59.

[12] Cambria, Erik, Devamanyu Hazarika, Soujanya Poria, Amir Hussain, and R. B. V. Subramanyam. "Benchmarking multimodal sentiment analysis." In International Conference on Computational Linguistics and Intelligent Text Processing, pp. 166-179. Springer, Cham, 2017.

[13] S. Nemati, R. Rohani, M. E. Basiri, M. Abdar, N. Y. Yen, and V. Makarenkov, "A Hybrid Latent Space Data Fusion Method for Multimodal Emotion Recognition," IEEE Access, vol. 7, 2019.

[14] N. M. Correa, T. Adali, Y.-O. Li, and V. D. Calhoun, "Canonical correlation analysis for data fusion and group inferences," IEEE signal processing magazine, vol. 27, no. 4, pp. 39–50, 2010.

[15] S. Poria, A. Hussain, and E. Cambria, Multimodal sentiment analysis. 2018.

[16] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, "Multimodal Sentiment Analysis: Addressing Key Issues and Setting Up the Baselines," IEEE Intelligent Systems, vol. 33, no. 6, pp. 17–25, 2018.

[17] K. Dashtipour, "Novel symbolic and sub-symbolic approaches for text based and multimodal sentiment analysis," 2019.

[18] M. E. Basiri, A. Kabiri, M. Abdar, W. K. Mashwani, N. Y. Yen, and J. C. Hung, "The effect of aggregation methods on sentiment classification in Persian reviews," Enterprise Information Systems, vol. 14, no. 9–10, 2020.

[19] K. Dashtipour, A. Hussain, Q. Zhou, A. Gelbukh, A. Y. A. Hawalah, and E. Cambria, "PerSent: A Freely Available Persian Sentiment Lexicon," 2016, pp. 310–320.

[20] M. E. Basiri, N. Ghasem-Aghaee, and A. R. Naghsh-nilchi, "Lexicon-based Sentiment Analysis in Persian," Current and Future Developments in Artificial Intelligence, p. 154, 2017.

[21] A. Amini, S. Karimi, and A. Shakery, "Cross-lingual Subjectivity Detection for Resource Lean Languages," 2019.

[22] R. Ghasemi, S. A. Ashrafi Asli, and S. Momtazi, "Deep Persian sentiment analysis: Cross-lingual training for low-resource languages," Journal of Information Science, 2020.

[23] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor Fusion Network for Multimodal Sentiment Analysis," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1103–1114.

[24] M. E. Basiri et al., "Improving Sentiment Polarity Detection through Target Identification," IEEE Transactions on Computational Social Systems, vol. 7, no. 1, 2020.

[25] M. E. Basiri, A. R. Naghsh-Nilchi, and N. Ghassem-Aghaee, "A framework for sentiment analysis in persian," *Open Transactions on Information Processing*, vol. 1, no. 3, pp. 1–14, 2014.

[26] E. Asgarian, M. Kahani, and S. Sharifi, "The Impact of Sentiment Features on the Sentiment Polarity Classification in Persian Reviews," Cognitive Computation, vol. 10, no. 1, pp. 117–135, Feb. 2018.

[27] S. M. Bagheri A., "Persian sentiment analyzer: A framework based on a novel feature selection method," International Journal of Artificial Intelligence, vol. 12,2,,115, no. pp. 115–129, 2014.

[28] M. B. Dastgheib, S. Koleini, and F. Rasti, "The application of Deep Learning in Persian Documents Sentiment Analysis," International Journal of Information Science and Management (IJISM), vol. 18, no. 1, pp. 1–15, 2020.

[29] B. Roshanfekr, S. Khadivi, and M. Rahmati, "Sentiment analysis using deep learning on Persian texts," in 2017 Iranian Conference on Electrical Engineering (ICEE), 2017, pp. 1503–1508.

[30] M. Heidari and P. Shamsinejad, "Producing An Instagram Dataset For Persian Language Sentiment Analysis Using Crowdsourcing Method."

[31] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[32] Kia Dashtipour, Mandar Gogate, Ahsan Adeel, Cosimo Ieracitano, Hadi Larijani, Amir Hussain. Exploiting Deep Learning for Persian Sentiment Analysis. arXiv:1808.05077. 2018 Aug 15.

[33] Zahra Bokaee Nezhad, Mohammad Ali Deihimi. A Combined Deep Learning Model for Persian Sentiment Analysis. IIUM Engineering Journal. 2019.

[34] Seyed Arad Ashrafi Asli, Behnam Sabeti, Zahra Majdabadi, Preni Golazizian, Reza Fahmi, Omid Momenzadeh. Optimizing Annotation Effort Using Active Learning Strategies: A Sentiment Analysis Case Study in Persian. Proceedings of the 12th Language Resources and Evaluation Conference. 2020.

[35] Javad PourMostafa Roshan Sharami, Parsa Abbasi Sarabestani, Seyed Abolghasem Mirroshandel. DeepSentiPers: Novel Deep Learning

Models Trained Over Proposed Augmented Persian Sentiment Corpus. arXiv:2004.05328. 2020.

[36] Nazanin Sabri, Ali Edalat, Behnam Bahrak. Sentiment Analysis of Persian-English Code-mixed Texts. arXiv:2102.12700. 2021.

[37] Kia Dashtipour, Cosimo Ieracitano, Francesco Carlo Morabito, Ali Raza. An Ensemble Based Classification Approach for Persian Sentiment Analysis. In book: Progresses in Artificial Intelligence and Neural Systems (pp.207-215). 2021.

[38] Nan Xu, Wenji Mao. A residual merged neutral network for multimodal sentiment analysis. 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA).

[39] Quoc-Tuan Truong, Hady Wirawan Lauw. VistaNet: Visual Aspect Attention Network for Multimodal Sentiment Analysis. Proceedings of the AAAI Conference on Artificial Intelligence 33:305-312. 2019.

[40] Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, Ajay Divakaran. Integrating Text and Image: Determining Multimodal Document Intent in Instagram Posts. arXiv:1904.09073. 2019.

[41] Mathieu Fortin, Brahim Chaib-Draa. Multimodal Sentiment Analysis: A Multitask Learning Approach. 8th International Conference on Pattern Recognition Applications and Methods. 2019.

[42] Akshi Kumar, Geetanjali Garg. Sentiment analysis of multimodal twitter data. Multimedia Tools and Applications. 2019.

[43] Ferda Ofli, Firoj Alam, Muhammad Imran. Analysis of Social Media Data using Multimodal Deep Learning for Disaster Response. arXiv:2004.11838. 2020.

[44] Manaswi, Navin Kumar. "Understanding and working with Keras." In Deep Learning with Applications Using Python, pp. 31-43. Apress, Berkeley, CA, 201

Aria Naseri Karimvand received his B.S. degree in software engineering from MJDKH university in 2019 and his M.S. from Shahrekord University in 2021. His research interest includes natural language processing, deep learning, and social media data mining.

Shahla Nemati was born in Shiraz, Iran in 1982. She received the B.S. degree in hardware engineering from Shiraz University, Shiraz, Iran, in 2005, the M.S. degree from Isfahan University of Technology, Isfahan, Iran, in 2008, and the Ph.D. degree in computer engineering from Isfahan University, Isfahan, Iran, in 2016. Since 2017, she has been an Assistant Professor with the Computer Engineering Department, Shahrekord University, Shahrekord, Iran. Her research interests include data fusion, affective computing, and data mining.

Reza Salehi Chegeni received his B.S. degree in software engineering from Lorestan university in 2018 and his M.S. from Shahrekord University in 2021. His research interest includes evolutionry algorithms, natural language processing, deep learning, and data mining.

Mohammad Ehsan Basiri received the B.S. degree in software engineering from Shiraz University, Shiraz, Iran, in 2006 and the M.S. and Ph.D. degrees in Artificial Intelligence from Isfahan University, Isfahan, Iran, in 2009 and 2014. Since 2014, he has been an Assistant Professor with the Computer Engineering Department, Shahrekord University, Shahrekord, Iran. He is the author of three books and more than 35 articles. His research interests include sentiment analysis, natural language processing, deep learning, and data mining