IJICTR
International Journal of Information & Communication Technology Research

Volume 16- Number 3 – 2024 (35 -44)

EmoRecBiGRU: Emotion Recognition in Persian Tweets with a Transformer-based Model, Enhanced by Bidirectional GRU

Faezeh Sarlakifar[1+] f.sarlakifar@mail.sbu.ac.ir
Morteza Mahdavi Mortazavi[1+] s.mahdavimortazavi@mail.sbu.ac.ir
Mehrnoush Shamsfard[1*] m-shams@sbu.ac.ir

[1]Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran

Received: 6 November 2023 – Revised: 14 February 2024 - Accepted: 20 April 2024

Abstract—**Emotion recognition in text is a fundamental aspect of natural language understanding, with significant applications in various domains such as mental health monitoring, customer feedback analysis, content recommendation systems, and chatbots. In this paper, we present a hybrid model for predicting the presence of six emotions: anger, disgust, fear, sadness, happiness, and surprise in Persian text. We also predict the primary emotion in the given text, including these six emotions and the "other" category. Our approach involves the utilization of XLM-RoBERTa, a pre-trained transformer-based language model, and fine-tuning it on two diverse datasets: EmoPars and ArmanEmo. Central to our approach is incorporating a single Bidirectional Gated Recurrent Unit (BiGRU), placed before the final fully connected layer. This strategic integration empowers our model to capture contextual dependencies more effectively, resulting in an improved F-score after adding this BiGRU layer. This enhanced model achieved a 2% improvement in the F-score metric on the ArmanEmo test set and a 7% improvement in the F-score metric for predicting the presence of six emotions on the final test set of the ParsiAzma Emotion Recognition competition.**

Keywords: Emotion Recognition (ER), Bidirectional Gated Recurrent Unit (BiGRU), Natural Language Processing (NLP), XLM-RoBERTa, Fine-tuning

Article type: Research Article

Publisher: ICT Research Institute

I. INTRODUCTION

Emotion recognition is a key part of natural language processing (NLP), a tech field that looks at different kinds of data, like text, to understand human emotions. This involves studying language, context, and behavior to determine how people feel. In this paper, we specifically focus on recognizing emotions in Persian text.

A. Importance of Emotion Recognition in NLP

Recognizing emotions from text is foundational for improving human-computer interactions, especially within chatbots. These automated conversational agents have become integral to various applications, and integrating emotion recognition capabilities can significantly enhance their effectiveness. Chatbots can tailor

[+] Equal Contributions
[*] Corresponding Author

[DOI: 10.61186/itrc.16.3.35]

responses to user sentiment by discerning emotions, fostering more personalized and engaging interactions.

However, emotion recognition is a challenging task. Several features, encompassed by the Component Process Model, express emotions, such as Stimulus, Bodily Symptoms, Subjective Feeling, Evaluation of Stimulus, expressions, and functional motivational aspects. These features present intricate patterns that pose difficulties for understanding and evaluation by computers.

A fundamental question arises: How can we define a categorical system of emotions? Various methods, such as Ekman's model [1] (including Happy, Sad, Fear, Anger, Surprise, and Disgust) and Robert Plutchik's Wheel of Emotions [2] (comprising joy, trust, fear, surprise, sadness, disgust, anger, and anticipation), have been proposed. In this work, we utilize Ekman's emotion categorization due to the availability of relevant datasets.

In the Persian language, sentiment analysis has garnered substantial attention, while emotion recognition has received relatively less focus, partly due to the scarcity of labeled datasets for emotion recognition tasks. Nonetheless, there are noteworthy works in Persian emotion recognition, including research conducted by Khotanlou et al. [3], another study by Abaskohi and colleagues [4], and the work of Mirzaee and colleagues [5], all of which have made valuable contributions. These studies are explored in detail in the related work section.

B. Related Work

Emotion recognition in the Persian language presents a distinctive array of challenges owing to the limited existing research, the intricacies inherent in analyzing the Persian language, and the scarcity of labeled datasets. Nevertheless, since the introduction of the EmoPars dataset [6], notable efforts have been made to enhance text-based Persian emotion recognition. Despite these endeavors, a considerable gap persists in achieving promising results in this domain.

Firstly, we review some important research in Emotion Recognition in general (English texts), then describe related works on Persian text emotion recognition.

Chowanda's research study compared the effectiveness of different machine learning models and a feed-forward neural network in recognizing emotions from the text. The study involves an exploration of various machine learning algorithms, including Naïve Bayes, Generalized Linear, Fast Large Margin, Artificial Neural Network (ANN), Decision Tree, Random Forest, and Support Vector Machine (SVM) [7]. This exploration involved analyzing 2302 feature sets, each containing 100-1000 features extracted from the text. The conclusion drawn from the results is that the Generalized Linear Model provides the best performance with an accuracy score of 0.92, recall of 0.902, precision of 0.902, and F1 score of 0.901.

The study conducted by Yakovenko addresses the challenge of multi-emotion sentiment classification in natural language processing (NLP). This work demonstrates the effectiveness of large-scale unsupervised language modeling combined with fine-tuning [8]. By training an attention-based Transformer network [9] on a substantial text dataset (Amazon reviews) and fine-tuning it on specific datasets. The results are acceptable for challenging emotion categories like Fear, Disgust, and Anger.

Related work in Persian emotion recognition has explored diverse methodologies to enhance performance. Noteworthy studies, such as Abaskohi's paper [4], have employed feature extraction techniques, focusing on emojis, hashtags, POS tags, and misspelled words, and addressing data imbalance issues. These studies have successfully fine-tuned transformer models, resulting in commendable outcomes.

Khotanlou’s paper [3], proposes a system for analyzing emotions in Persian texts, combining cognitive features and a deep neural network, specifically a gated recurrent unit (GRU) [10]. Using a dataset of 23,000 labeled Persian documents, the approach incorporates emotional constructions, keywords, and POS, along with Word2Vec for text embedding.

In Mirzaee’s paper [5], whose public dataset was utilized in our work, the authors endeavored to create a deep learning model using transfer-learning and preprocessing techniques on text. They fine-tuned transformer models, such as BERT and its family, after processing the text and achieved favorable baseline results for their new dataset (ArmanEmo Dataset).

Some models, like “XLM-T” adopted transformer models, fine-tuning them on a large volume of informal texts from platforms like Twitter and Instagram to enhance the model's understanding of informal language. Subsequently, these models were fine-tuned on emotion recognition datasets, predominantly comprised of informal texts from Twitter [11].

Other approaches have relied on classical machine learning methods, lexicon-based approaches, and phrase-based methods, which have laid the foundation for the field [12]. However, as we advance into the future, end-to-

[DOI: 10.61186/itrc.16.3.35]

end deep learning techniques are gaining prominence.

Beyond the Persian language, current efforts are increasingly focused on creating end-to-end pipelines to comprehend features in text and detect emotions. For instance, in Kumar's paper [13], the authors aimed to create strong pipelines by improving how text is understood and effectively detecting emotions. They utilized a bidirectional encoder representation transformer as a powerful embedding module, combined GRU and CNN blocks, and finished with a feed-forward network for the classifier. These efforts focus on integrating different neural network components and experimenting to build a robust end-to-end pipeline.

C. Motivation

Our motivation for conducting this research was sparked by the “ParsiAzma competition” [14]. This national NLP competition comprises four challenges in the analysis of social media text:

1. Emotion Recognition
2. Sentiment Analysis
3. Fact-Checking
4. Stance Detection

This competition constitutes four primary stages, and the results of the fourth stage determine the final competition rankings. In our chosen challenge, “Emotion Recognition,” our proposed model achieved second place. Additionally, there was a fifth stage dedicated to the improvement phase, allowing us to test our enhanced models following the conclusion of the primary competition stages.

Deep learning approaches enable models to automatically learn features from text, enhancing their ability to better understand the nuances of emotional content. Our work aligns with this approach, aiming to enhance model performance through the integration of attention mechanisms and recurrent neural networks (RNN).

Our inspiration for incorporating a GRU layer into XLM-RoBERTa stems from Chin Poo Lee's recent paper [15], We hypothesized that this additional layer could enhance results in the Persian text emotion recognition task, and the achieved outcomes align with our expectations. This paper will detail our model's architecture and methodology, experimental findings, and conclusion, along with future work considerations.

II. METHODOLOGY

A. Datasets

In our research, two primary emotion datasets are utilized: EmoPars [6] and ArmanEmo [5].

1) EmoPars: EmoPars is a substantial dataset comprising 30,000 sentences sourced from Twitter. Each sentence in this dataset is associated with specific emotions, including Anger, Fear, Happiness, Hatred, Sadness, and Wonder. What sets EmoPars apart is its unique feature where each emotion in a sentence is assigned a score ranging from 0 to 5. This score signifies the intensity of the corresponding emotion in the text. In our work, we aimed to predict the presence or absence of specific emotions in a sentence. To do this, we categorized each emotion pair with "0" indicating absence, and "1" indicating presence.

2) ArmanEmo: The ArmanEmo dataset comprises 7,000 sentences collected from a variety of sources, including Twitter, Instagram, and comments on Digikala – The greatest online shop in Iran. The objective behind this diverse collection was to create a dataset that offers a broader representation of emotions. However, it's worth noting that this dataset, while more comprehensive in terms of sources, contains fewer samples compared to EmoPars. Consequently, it may exhibit more noise and lower overall integrity. Nevertheless, this dataset is valuable for its coverage of 7 emotion classes: Sad, Hate, Angry, Fear, Happy, Surprise, and "Other." Notably, these labels align with the test data of the ParsiAzma challenge, which enabled us to predict the primary emotion of a sentence effectively. Additionally, this dataset is published in split train and test sets, enabling us to compare our results on its test set with the work of others.

3) Preprocessing in detail

In our study, a preprocessing pipeline is implemented that closely follows the methodologies outlined in Mirzaee’s paper [5], with a few adjustments. Our preprocessing steps aimed to ensure that the textual data was in an optimal format for inputting into our models. The steps involved in this process were as follows:

1) Normalization and Full-Cleaning with Dadmatools [16]: Dadmatools cleaning includes the following steps: unify_chars, refine_punc_spacing, remove_extra_space, remove_puncs, remove_html, remove numbers, and remove URLs.

2) Handling of Repeated Letters: Informal texts often include Persian words with repeated letters for emphasis (e.g., "الووو", "عاااللیییی", "خیلللییییی"). We corrected these non-standard

[DOI: 10.61186/itrc.16.3.35]

spellings using the Hazm normalizer. This preprocessing step was done similarly to the approach described in the ArmanEmo paper.

3) Removal of Non-Persian Characters: Any non-Persian characters, including Arabic Diacritics, English Characters, and so on, were removed. Before doing so, we extracted emojis as they are valuable features in our input.

3.1) Removal of Arabic Diacritics: Persian words may be written with or without Arabic diacritics. To standardize our text, we removed all Arabic diacritics.

3.2) Removal of English Characters: we eliminated any English characters to ensure that our text is exclusively in Persian. Although our employed model (XLM-RoBERTa) was a multilingual model capable of recognizing English and Arabic characters, following the ArmanEmo preprocessing methods, which remove all non-Persian characters, led to better results.

4) Removal of Persian Numeric Characters: We eliminated Persian numeric characters from the text.

5) Handling of Hashtags: Instead of removing hashtag signs, we preserved the information within the hashtags, retaining them and feeding them to the SentencePiece tokenizer. This approach was adopted as hashtags can provide significant text features.

6) Handling of Emojis: We extracted emojis and, along with the extracted hashtags, fed them to the tokenizer as another impactful feature. Emojis can quickly determine the target emotion just by themselves.

Previously, we mentioned that our labels were not in binary classification form, so we converted them into binary form. Another challenge was the difficulty in ensuring the correctness of this conversion due to the data annotation policy. The labels represent scores in the range of 0 to 5. thus, threshold setting was necessary. This issue is also discussed in the Abaskoh's paper [4].

For the EmoPars dataset [6], which lacked binary classification labels for emotions, our approach involved the following steps:

Normalization of Scores: We normalized the emotion scores, originally ranging from 0 to 5, to fit within the range of 0 to 1.

Threshold Assignment: We set a threshold for each emotion, such as 0.4, to determine an emotion's presence (label 1) or absence (label 0) in the input. Using a trial-and-error approach, it was found that the threshold range between 0.35 and 0.5 proved effective for our task.

These preprocessing steps allowed us to transform raw text into a standardized format suitable for input to our models. Importantly, we consistently applied these modules to both Arman and EmoPars datasets, in both the training and test data.

B. Proposed model

In this section, we present our final architecture. The foundation of our model is rooted in transformer language models, serving as the embedding component.

We experimented with various models, including Recurrent Neural Network (RNN) models such as Long Short-Term Memory (LSTM) [17], Multilingual BERT, Distill BERT, ParsBERT, and XLM-RoBERTa. By tracking and enhancing results, we found that XLM-RoBERTa outperformed Bert and ParsBERT due to its special training configuration. Consequently, we established XLM-RoBERTa as the embedding and core of our architecture. We then explored different methods for efficient fine-tuning, incorporating various neural networks and monitoring results.

Our model integrates two XLM-RoBERTa models with an additional Bidirectional GRU layer. Fine-tuning is conducted on distinct datasets: The EmoPars dataset and the ArmanEmo dataset.

RoBERTa: Short for "A Robustly Optimized BERT Pre-training Approach," is a language model pre-training technique optimized for natural language understanding tasks [18].

XLM-RoBERTa: A multilingual variant of RoBERTa pre-trained on a vast dataset encompassing 100 languages, including the Persian language [19].

GRU: Abbreviation for "Gated Recurrent Units," [10] a type of recurrent neural network (RNN) architecture used for sequential data processing with efficient long-range dependency capture.

Combining GRU with RoBERTa embedding, as tested in Chin Poo Lee's recent paper [15], showed promise for sentiment analysis benchmarks and, in our estimation, could be effective for emotion recognition. We modified the number of GRU layers and utilized XLM-RoBERTa for its multilingual feature.

Our model independently addresses two core tasks: predicting all emotions and predicting the primary emotion. In simpler terms, "all emotions prediction" means determining the presence (1) or absence (0) of each of the six different emotions in a given text, while "primary emotion detection" refers to finding the most dominating emotion in the text.

[DOI: 10.61186/itrc.16.3.35]

The final hybrid model predicts all emotions present in the given text using an XLM-RoBERTa model with an additional Bidirectional GRU layer, that is fine-tuned on the entire EmoPars dataset. On the other hand, it predicts the primary emotion using only the XLM-RoBERTa model, fine-tuned on the ArmanEmo dataset.

Due to the limitations of the BERT family, especially ParsBERT [20], in capturing robust contextual information, we opted for XLM-RoBERTa for the primary emotion part as well. Adding GRU to XLM-RoBERTa on the ArmanEmo dataset resulted in overfitting due to the relatively scarce training data. (Given the smaller number of data samples in ArmanEmo compared to EmoPars, which has 30k samples). Therefore, we solely fine-tuned XLM-RoBERTa on the ArmanEmo dataset. Throughout, we employed two XLM-RoBERTa models: XLM-RoBERTa (base version) and XLM-RoBERTa (large version). While both were utilized during our experiments, XLM-RoBERTa (large version) was selected for our final model. Additionally, we tested XLM-RoBERTa-base's performance on the ParsiAzma fourth stage test set.

C. *Evaluation metrics*

Precision (P) for each class (c):

$$P_C = \frac{TP_C}{TP_C + FP_C} \quad (1)$$

Recall (R) for each class (c):

$$R_C = \frac{TP_C}{TP_C + FN_C} \quad (2)$$

F1 score (F1) for each class (c):

$$F1_C = \frac{P_C \times R_C}{\frac{P_C + R_C}{2}} \quad (3)$$

Macro-Average F1 Score:

$$\text{Macro\_F1} = \frac{1}{N}\sum_{c=1}^{N} F1_C \quad (4)$$

Macro Precision:

$$\text{Precision\_Macro} = \frac{1}{N}\sum_{c=1}^{N} P_C \quad (5)$$

Macro Recall:

$$\text{Recall\_Macro} = \frac{1}{N}\sum_{c=1}^{N} R_C \quad (6)$$

Where:

- TP_c is the number of true positives for class c.
- FP_c is the number of false positives for class c.
- FN_c is the number of false negatives for class c.
- N is the total number of classes.

In summary, our final hybrid model predicts all emotions presence using an XLM-RoBERTa (large version) with a Bidirectional GRU layer, fine-tuned on the entire EmoPars dataset, in addition to employing the same model without GRU is fine-tuned on the entire ArmanEmo dataset for predicting the primary emotion. The overall model architecture is shown in Figure 1.

III. Experiments and Results

We assessed the performance of our proposed emotion recognition model using cross-validation on EmoPars. The dataset comprises 30,000 labeled instances, encompassing six emotions: anger, sadness, fear, wonder, happiness, and hatred.

Employing a 5-fold cross-validation, we divided the dataset into five equal-sized subsets. The model was trained on four subsets and evaluated on the remaining one, repeating this process five times to ensure each subset served as both training and testing data. We measured our model's performance using precision, recall, F-score, and accuracy as evaluation metrics. The obtained results are presented in Table 1.

Furthermore, we explored the use of "ParsBERT + GRU layer" for primary emotion prediction in the ArmanEmo test set, noting a 2% improvement in F-score compared to the results reported in Abaskohi's paper [4] on the ArmanEmo test set. In their paper, they defined the ArmanEmo dataset and fine-tuned the ParsBERT model on the ArmanEmo's train set, testing it on the ArmanEmo's test set. We followed this approach and added a GRU layer to assess its impact, and the results highlighted its effectiveness in enhancing the F-score.

Based on the statistics in Table 1, our final aggregated result on the EmoPars test sets was 0.62 macro-F-score. While achieving these results, we encountered challenges, that will be explained in the Discussion section. Additionally, the results of our proposed model on the ParsiAzma final test set are shown in Table 2.

Comparing the results on the ParsiAzma test set with five-fold cross-validation on EmoPars, we observed that results on the ParsiAzma final test set were even better than the results of the five-fold cross-validation on EmoPars dataset. This difference can be attributed to training our model on the entire EmoPars dataset (30,000 instances) for ParsiAzma, whereas we used 24,000 instances for training and 6,000 for testing in each fold of the cross-validation.

Despite the noisy nature of the EmoPars dataset, the results were comparable to ParsiAzma, suggesting a similarity between the two datasets.

[DOI: 10.61186/itrc.16.3.35]

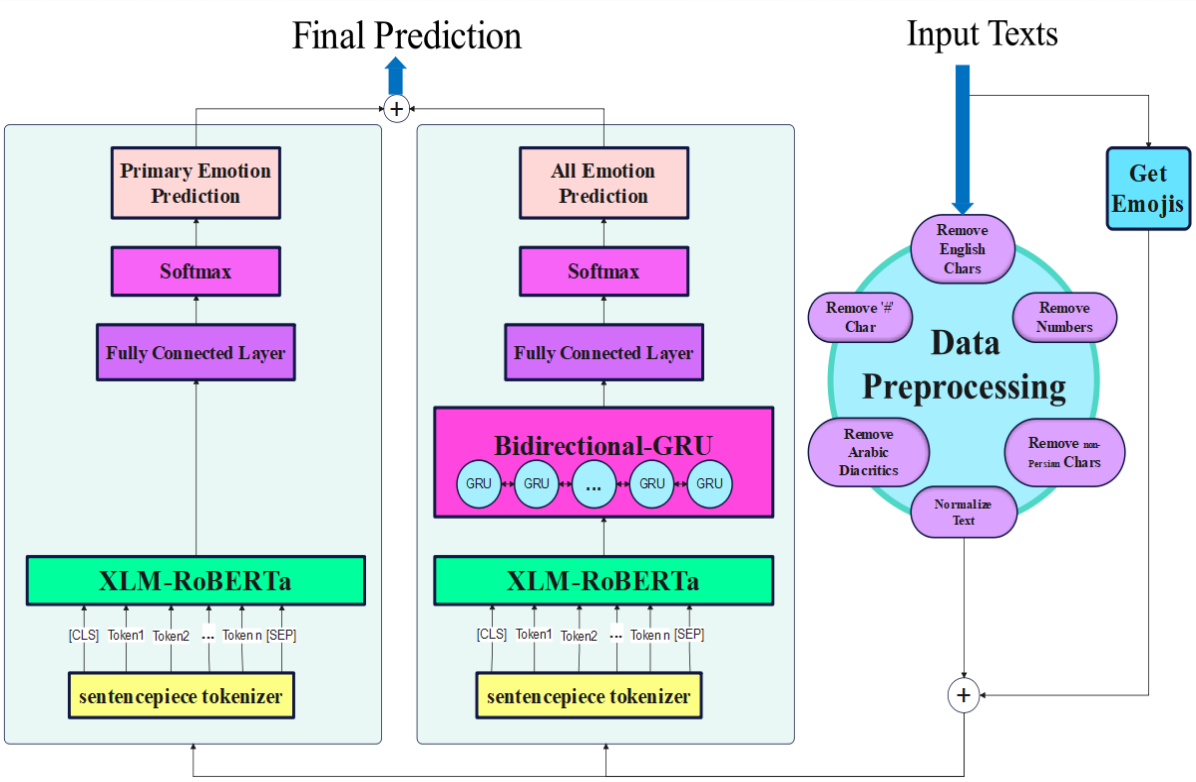

Figure 1. Block diagram of the proposed Architecture for Persian text emotion recognition.

TABLE I. RESULTS OF ALL EMOTION PREDICTION USING 5-FOLD CROSS-VALIDATION ON EMOPARS DATASET

Set	AE Accuracy	AE weighted-F-score	AE macro-F-score
1	0.85	0.80	0.68
2	0.81	0.73	0.60
3	0.82	0.76	0.64
4	0.77	0.71	0.56
5	0.82	0.74	0.61
Average	**0.81**	**0.75**	**0.62**

Our work faced a new challenge related to primary emotion prediction. In the EmoPars dataset, the "other" label was absent, leading to challenges in prediction. To address this, we utilized the ArmanEmo dataset, which had seven labels, including "other," making it compatible with the ParsiAzma challenge test set labels. Initially, we employed the ParsBERT [20] model for embedding and made incremental improvements over Arman's baseline. We achieved a 1% improvement in F-macro by modifying the training configuration (changing the learning rate and batch size). The final model, XLM-RoBERTa (large version), was selected for its robust ability to extract relations between tokens. We trained this model by fine-tuning and adding one Bidirectional GRU layer. While adding Bidirectional GRU performed better on EmoPars (30,000 samples), it did not work well on the ArmanEmo Test set due to a lack of training data, causing overfitting. Therefore, we only fine-tuned XLM-RoBERTa (large version) and We achieved a score of 0.57 in F-score-macro. Our results in the fourth stage of ParsiAzma are presented in Table 3.

TABLE II. RESULTS ON THE PARSIAZMA EMOTION RECOGNITION CHALLENGE, FIFTH STAGE. (FIFTH STAGE IS THE IMPROVEMENT PHASE, WHICH IS THE ADDITIONAL STAGE AFTER THE FINAL PRIMARY COMPETITION STAGE AND THE DETERMINATION OF RANKINGS)

Team	AV fscore[1]	AE fscore[2]	AE recall[3]	AE precision[4]	PE fscore[5]	PE recall[6]	PE precision[7]

[DOI: 10.61186/itrc.16.3.35]

Our Proposed Model Fine-tuning **XLM-RoBERTa (large) + Bidirectional GRU** on EmoPars for AE prediction, Fine-tuning **XLM-RoBERTa (large)** on ArmanEmo for PE prediction	**0.51**	**0.63**	**0.77**	**0.68**	**0.38**	**0.46**	**0.56**

TABLE III. RESULTS ON THE PARSIAZMA EMOTION RECOGNITION CHALLENGE, THIRD STAGE.

Team	AV fscore	AE fscore	AE recall	AE precision	PE fscore	PE recall	PE precision
First Rank	0.62	0.66	0.62	0.73	0.58	0.67	0.58
Second Rank: Our Proposed Model	**0.47**	**0.57**	**0.84**	**0.47**	**0.38**	**0.46**	**0.56**
Third Rank	0.46	0.54	0.77	0.43	0.39	0.47	0.43
Fourth Rank	0.31	0.33	0.29	0.46	0.29	0.35	0.31

As explained in the next section, we encountered challenges while predicting primary emotions on the ParsiAzma test sets. In the fourth stage (the final primary stage of the competition), we tested various models, and the results are reported in Table 4. In this stage, our additional GRU layer was unidirectional. Our final proposed model contains a Bidirectional GRU layer. This update from a unidirectional to a bidirectional GRU layer led to a 1% improvement in the macro-F-score for all emotion predictions. A bidirectional GRU can capture dependencies better than a unidirectional GRU layer; therefore, this improvement in F-score was expected. The results of adding a bidirectional GRU are shown in Table 2. The results of using a unidirectional GRU layer are shown in Table 4. These results are also displayed as a bar chart in Fig. 2.

TABLE IV. RESULTS ON THE PARSIAZMA EMOTION RECOGNITION CHALLENGE FINAL TEST SET (FOURTH STAGE)

Model	AV fscore	AE fscore	AE recall	AE precision	PE fscore	PE recall	PE precision
"Model 1" Fine-tuning **XLM-RoBERTa (base)** for AE prediction Fine-tuning **ParsBERT [20]** on ArmanEmo for PE prediction	0.45	0.62	0.73	0.58	0.28	0.36	0.43
"Model 2" Fine-tuning **XLM-RoBERTa (base) + GRU** on EmoPars for AE prediction Fine-tuning **XLM-RoBERTa (base)** on ArmanEmo for PE prediction	0.46	0.59	0.86	0.49	0.33	0.39	0.47
"Model 3" Fine-tuning **XLM-RoBERTa (large) + GRU** on EmoPars for AE prediction Fine-tuning **XLM-RoBERTa (large) + GRU** on ArmanEmo for PE prediction	0.49	0.62	0.73	0.58	0.35	0.41	0.47
"Model 4" Fine-tuning **XLM-RoBERTa (large) + GRU** on EmoPars for AE prediction Fine-tuning **XLM-RoBERTa (large)** on ArmanEmo for PE prediction	**0.50**	**0.62**	**0.73**	**0.58**	**0.37**	**0.42**	**0.49**

[1]Macro-Average F1 Score
[2]All-Emotion Macro F1 Score
[3]All-Emotion Macro Recall
[4]All-Emotion Macro Precision
[5]Primary-Emotion Macro F1 Score
[6]Primary-Emotion Macro Recall
[7]Primary-Emotion Macro Precision

[DOI: 10.61186/itrc.16.3.35]

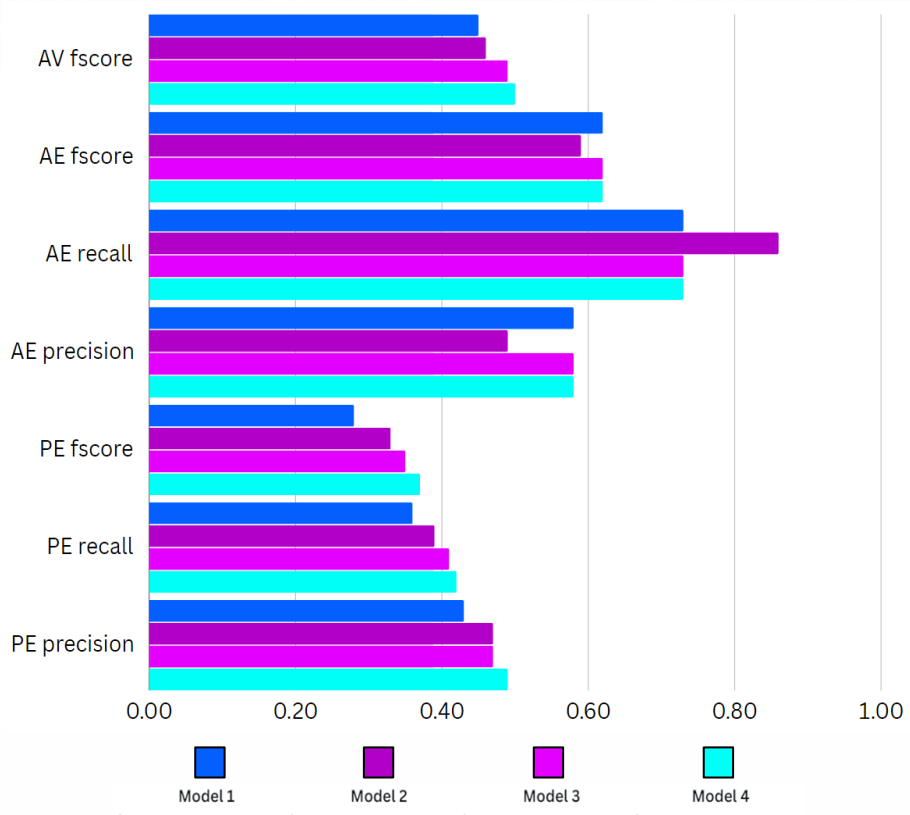

Figure 2. A chart displaying the results of the ParsiAzma Emotion Recognition challenge test set (fourth stage) for our various models.

IV. Discussion

The diverse sources used for the ArmanEmo dataset, for instance, using Digikala—The e-commerce platform digikala.com [21] provides a broad selection of consumer goods, encompassing electronics, groceries, personal care items, and digital products—comments alongside Twitter and Instagram while mean they have different contexts, resulting in a distinct distribution in comparison with the ParsiAzma [14] test sets, introducing noise. Utilizing various resources for creating a robust benchmark dataset is commendable and can be good for a better comprehension of the model, our specific concern arises from the fact that the ArmanEmo train set consists of only 6,000 instances. With such a small dataset, incorporating diverse resources leads to noise and model overfitting. Consequently, when attempting to evaluate the model on integrated data, such as data sourced exclusively from Twitter, suboptimal results are obtained. If different resources are to be used for creating a dataset, it must be large enough to encompass a variety of contexts, and the noise can be ignored due to its large size. It is noteworthy that deploying ArmanEmo and evaluating it on its test set consistently provides positive outcomes. However, when applied to a dissimilar dataset with a distinct distribution (like tweets), unsatisfactory results are obtained. To address these challenges and enhance the performance of our model, we propose mitigating data bias, implementing more robust handcrafted feature engineering, and exploring self-supervised training methodologies on substantial amounts of informative data, such as that derived from Twitter. Such measures hold the potential to improve the model's performance.

Our experiments show that in the “All Emotion” task, replacing the unidirectional GRU layer with the bidirectional GRU can enhance the model's performance across all evaluation metrics. This change was caused by a 10% improvement in precision, a 4% improvement in recall, and a 1% improvement in the average macro-F-score.

The above improvement is due to the bidirectional GRU layer processing the input sequence both forward and backward. This means that at each time step, the hidden state of the GRU unit is influenced by both past and future tokens in the input sequence. In emotion recognition tasks, understanding the context of a given word or phrase is crucial for accurately determining the emotion expressed. By incorporating information from preceding and succeeding words, bidirectional GRU layers can capture a more comprehensive contextual understanding, leading to improved performance.

Furthermore, bidirectional GRUs are better equipped to capture long-range dependencies in the input sequence compared to unidirectional GRUs. Emotions in text often depend on complex

[DOI: 10.61186/itrc.16.3.35]

relationships between words or phrases that may occur far apart in the text. Bidirectional processing allows the model to better capture these dependencies, which can lead to better predictions.

In our research study, we aimed to assess the overall model performance across all emotions. Therefore, we found the macro-F-score to be a more suitable evaluation metric. The macro-average F-score computes the F1 score for each emotion class independently and then averages them. This approach ensures that each emotion class contributes equally to the final score, irrespective of its frequency in the dataset.

On the contrary, the weighted F-score calculates the F1 score for each class but takes into account the support (i.e., the number of true instances) of each class. As a result, emotions with higher instances carry more weight in the final score, reflecting their significance in the dataset.

Given our objective of assessing the model's overall performance across all emotion classes, the macro-average F-score provided a clear and unbiased measure by averaging the F1 scores across all classes.

V. Conclusion

In summary, our research represents a hybrid model for recognizing emotions in Persian text. We utilized two distinct datasets, EmoPars and ArmanEmo, and used deep learning techniques to introduce a model that can effectively predict emotions in Persian text. This model predicts the “Primary Emotion” by classifying input text into seven categories, including six specific emotions and the “other” category. Additionally, it performs the "All Emotion" task, determining whether each emotion is present in the given text or not.

Our results support the hypothesis that similar to English texts, incorporating a GRU layer—whether unidirectional or bidirectional—can enhance the performance of emotion recognition in Persian texts, particularly in the “All Emotion” task. However, challenges were encountered in the “Primary Emotion” task due to limitations inherent in the available Persian text datasets.

A. Future Work

In our upcoming work, we plan to enhance our model's capabilities to predict both sentiment and emotion, offering a more comprehensive understanding of textual content. This generalized model can be more profitable. We can further improve our model's performance in all emotion predictions by incorporating insights from primary emotion predictions, and vice versa. This approach would involve creating a hybrid model, integrating two models that are not entirely independent. Moreover, the potential solutions to address challenges related to data limitations in primary emotion prediction, as introduced in the discussion section, can be considered good options for our future work.

Acknowledgment

Thanks to the organizers of the ParsiAzma Competition for providing the platform and opportunity to conduct this research. Their dedication to advancing the field of emotion recognition has been a driving force behind our project.

Conflict of Interest

We have no conflicts of interest to disclose.

References

[1] Ekman, P. An argument for basic emotions. Cogn. Emot. 6(3–4), 169–200 (1992)

[2] Plutchik, R. The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and eprovide tools for clinical practice. American Scientist, 89(4), 344-350 (2001).

[3] S. S. Sadeghi, H. Khotanlou, and M. Rasekh Mahand, “Automatic Persian Text Emotion Detection using Cognitive Linguistic and Deep Learning,” in Journal of Artificial Intelligence and Data Mining (JAIDM), 2020, pp 169-179

[4] A. Abaskohi, N. Sabri, and B. Bahrak, “Persian Emotion Detection using ParsBERT and Imbalanced Data Handling Approaches,” in ArXiv, 2022.

[5] H. Mirzaee, J. Peymanfard, H. H. Moshtaghin, and H. Zeinali, ‘ArmanEmo: A Persian Dataset for Text-based Emotion Detection’, ArXiv, vol. abs/2207.11808, 2022.

[6] N. Sabri, R. Akhavan, B. Bahrak, “EmoPars: A Collection of 30K Emotion-Annotated Persian Social Media Texts,” in ACL Anthology: RANLP 2021, pp 167–173.

[7] A. Chowanda, R. Sutoyo, Meiliana, S. Tanachutiwat, “Exploring Text-based Emotions Recognition Machine Learning Techniques on Social Media Conversation,” in Procedia Computer Science, 2021, pp 821-828.

[8] N. Kant, R. Puri, N. Yakovenko, and B. Catanzaro, ‘Practical Text Classification With Large Pre-Trained Language Models’, ArXiv, vol. abs/1812.01207, 2018.

[9] A. Vaswani et al., ‘Attention is all you need’, in Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 2017, pp. 6000–6010.

[10] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, ‘On the Properties of Neural Machine Translation: Encoder-Decoder Approaches’, in Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics, and Structure in Statistical Translation, 2014, pp. 103–111.

[11] F. Barbieri, L. Espinosa Anke, and J. Camacho-Collados, ‘XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond’, in Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 258–266.

[12] Ali Khosravi, M. Kelarestaghi, M. Purmohammad, “Emotion Detection in Persian Text; A Machine Learning Model,” in Biannual Journal of Contemporary Psychology, 2019, pp 42-48.

[13] P. Kumar and B. Raman, “A BERT based dual-channel explainable text emotion recognition system”, in Neural Networks, 2022, pp 392-407.

[14] ParsiAzma National NLP Competition, 2023, Final Results, accessed on Feb. 2024.

[DOI: 10.61186/itrc.16.3.35]

[15] K. L. Tan, C. P. Lee, and K. M. Lim, "RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment Analysis," in MDPI Applied Sciences, 2023.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, 1997.

[17] R. Etezadi, M. Karrabi, N. Zare, M. B. Sajadi, M. T. Pilehvar, "DadmaTools: Natural Language Processing Toolkit for Persian Language," in ACL Anthology, 2022., pp 124-130.

[18] L. Zhuang, L. Wayne, S. Ya, and Z. Jun, 'A Robustly Optimized BERT Pre-training Approach with Post-training', in Proceedings of the 20th Chinese National Conference on Computational Linguistics, 2021, pp. 1218–1227.

[19] A. Conneau et al., 'Unsupervised Cross-lingual Representation Learning at Scale', in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.

[20] M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri, 'ParsBERT: Transformer-based Model for Persian Language Understanding', Neural Process. Lett., vol. 53, no. 6, pp. 3831–3847, Dec. 2021.

[21] https://about.digikala.com/en/about-us accessed on Feb. 2024.

Morteza Mahdavi Mortazavi is a B.Sc. student in Computer Engineering at Shahid Beheshti University, starting in 2021. His academic journey is deeply rooted in a passion for machine learning, with a particular focus on Natural Language Processing (NLP). His research interests center around Sentiment Analysis, Emotion Detection and the Development and Optimization of Large Language Models (LLMs).

Faezeh Sarlakifar is a B.Sc. student in Computer Engineering at Shahid Beheshti University. She started her B.Sc. in 2019. Her research interests are focused on applied deep learning in Bioinformatics, Natural Language Processing, and Healthcare. She is passionate about working with Large Language Models (LLMs) and exploring their potential to enhance Data-Driven Decision-Making and Innovation in these fields.

Dr. Mehrnoush Shamsfard has received her B.Sc. and MSc both on Computer Software Engineering from Sharif University of Technology and her Ph.D. in Computer Engineering-Artificial Intelligence from AmirKabir University of Technology, Tehran, Iran. She has been with Shahid Beheshti University from 2004. She is currently Associate Professor of Faculty of computer science and engineering, and also the head of NLP research Laboratory of this faculty. Her main fields of interest are Natural Language Processing, LLMs, Developing Intelligent Assistants, Evaluating NLP Products and Resources and Knowledge Engineering (Ontologies and Knowledge Graphs).

[DOI: 10.61186/itrc.16.3.35]