IJICTR
International Journal of Information & Communication Technology Research

Volume 16- Number 3 – 2024 (9 -18)

A Comparative Study of BERT-X for Sentiment Analysis and Stance Detection in Persian Social Media

Mohamad Sobhi
Department of Computer Engineering
Amirkabir University of Technology
Tehran, Iran
sobhi@aut.ac.ir

Alireza Mazochi
Department of Computer Engineering
Amirkabir University of Technology
Tehran, Iran
mazochi@aut.ac.ir

Hossein Zeinali[*+]
Department of Computer Engineering
Amirkabir University of Technology
Tehran, Iran
hzeinali@aut.ac.ir

Received: 16 December 2023 – Revised: 25 February 2024 - Accepted: 13 April 2024

Abstract—**BERT-based models have gained popularity for addressing various NLP tasks, yet the optimal utilization of knowledge embedded in distinct layers of BERT remains an open question. In this paper, we introduce and compare diverse architectures that integrate the hidden layers of BERT for text classification tasks, with a specific focus on Persian social media. We conduct sentiment analysis and stance detection on Persian tweet datasets. This work represents the first investigation into the impact of various neural network architectures on combinations of BERT hidden layers for Persian text classification. The experimental results demonstrate that our proposed approaches can outperform the vanilla BERT that utilizes an MLP classifier on top of the corresponding output of the CLS token in terms of performance and generalization.**

Keywords: BERT, Persian Text Classification, Social Media, Sentiment Analysis, Stance Detection, CNN, LSTM.

Article type: Research Article

Publisher: ICT Research Institute

[DOI: 10.61186/itrc.16.3.9]

I. INTRODUCTION

Social media analysis seeks to extract and interpret information from social media platforms, encompassing a broad spectrum of data, including opinions, sentiments, emotions, preferences, stances, and trends. This type of analysis offers valuable insights into various aspects of human behavior and society. In this context, we have chosen sentiment analysis and stance detection as two pivotal tasks within social media analysis.

* Corresponding Author
+ Equal Contributions. The order of the first two authors is random.

Sentiment analysis involves identifying and extracting subjective opinions and emotions expressed in texts, classifying them as positive, negative, or neutral. Stance detection, on the other hand, focuses on extracting a user's response to a text written by another user and is integral to approaches for detecting fake news.

This article specifically concentrates on the Persian language. BERT[1] [1] and its extensions stand out as the state-of-the-art models for Persian text classification. Various monolingual and multilingual versions of BERT have been developed for Persian, and for this work, we employ XLM-RoBERTa, a multilingual model based on the BERT architecture, which we will refer to as BERT for simplicity throughout this paper. While achieving good results can be as simple as applying a Multi-Layer Perceptron (MLP) solely on the [CLS] token, our approach involves using a more complex network on all tokens, particularly from some of the last layers, to enhance accuracy.

The majority of existing works utilize BERT with a default pooling layer, relying solely on the final hidden state of the [CLS] token and one or more fully connected layers. Some studies have incorporated a Convolutional Neural Network (CNN) on BERT output, referred to as BERT-CNN in literature [2-4]. Generally, we categorize these models as BERT-X, where X represents the auxiliary network name. However, few comprehensive comparative studies of BERT-X models for social media analysis have been conducted in Persian, and the absence of available code poses an additional challenge.

To address these challenges, we propose several BERT-X models, including BERT-MLP, BERT-LSTM, BERT-CNN, BERT-WeightingCLS, and BERT-ClsAverageMax, drawing inspiration from models in other languages. These models are systematically compared in two Persian social media analyses: sentiment analysis and stance detection. Moreover, these models can be easily adapted for other Persian text classification tasks. The results indicate that BERT-LSTM achieves the highest performance for sentiment analysis with an F1-macro score of 64.56% and an F1-micro score of 73.00%. Similarly, BERT-WeightingCLS attains the best results for stance detection with an F1-macro score of 66.30% and an F1-micro score of 71.58%.

The subsequent sections of this article are organized as follows: Section 2 provides a literature review covering two aspects of our work—BERT-X and Persian social analysis. Section 3 outlines our proposed models, offering details on task definition and model architectures. Section 4 presents our experiments and analysis, while Section 6 concludes the article and suggests directions for future research.

II. Related Works

This section provides a review of the relevant literature related to our research, focusing on two main aspects: research method and research domain. Our research methodology revolves around BERT, and thus, the initial subsection offers a comprehensive overview of models with different heads. The research domain specifically explores Persian social media analysis, encompassing Persian sentiment analysis and Persian stance detection. The second subsection summarizes Persian literary works related to sentiment analysis and stance detection.

A. BERT-X

BERT remains an efficient and state-of-the-art approach for various natural language understanding tasks. Originally, the [CLS] token of BERT was utilized for sentence classification, often employing a simple MLP on the [CLS] vector for this purpose. Our BERT-MLP is part of this model group. Additionally, we introduce BERT-WeightingCLS, which utilizes a dynamic average of the last four layers [CLS] embeddings before applying an MLP. Recent works suggest using all embedding tokens, not just [CLS], involving some last layers rather than just the final one, and employing a more complex network on BERT. Lehecka et al. [5] employed a time-distributed feedforward on all text tokens (excluding [CLS] token), followed by max-pooling and average-pooling on the time-distributed feedforward outputs. Our BERT-ClsAverageMax is inspired by this work, with the distinction of using concatenation instead of summation to enhance network capacity.

Several studies employ a combined architecture of BERT and CNN. In some of these instances, BERT is utilized for the initial representation, with this representation then inputted into a CNN for further representation. Subsequently, these two representations are processed in a network for the final prediction. Zheng et al. introduced one of the early architectures for classification using BERT and CNN. In this study, BERT is employed for the representation of each token in the initial step, followed by the application of kx1 convolution to text tokens (excluding [CLS] and [SEP]) to extract local information. The [CLS] embedding and the CNN outputs are fed into another Transformer encoder for the final prediction [6].

Wan et al. [7] investigated financial causal sentence recognition using a model very similar to Zeng et al.'s [6] work. Dong et al., in their examination of commodity sentiment analysis, utilized a BERT-CNN model. Their model employs the [CLS] token of BERT as the initial sentence representation to extract global features. This representation is then fed into a simple CNN network consisting of convolution and pooling layers for local feature extraction. The concatenation of the BERT output and CNN output is processed with a single, straightforward classification layer [2].

Jia proposed a framework based on BERT, CNN, and attention mechanisms for the sentiment classification of Chinese microblogs. The [CLS] token's output serves as the text embedding, while the embeddings of other tokens are input into a CNN. The CNN incorporates a convolutional layer, top-k-average pooling, attention pooling, and dense layers. The outputs of BERT and CNN feed into another attention layer [8].

[1] Bidirectional Encoder Representations from Transformers

[DOI: 10.61186/itrc.16.3.9]

Some researchers prefer a pipeline involving BERT and CNN, wherein BERT processes inputs, and the processed output is then fed to a CNN. Some works incorporate an additional network to process the CNN outputs. Kaur et al. developed a BERT-CNN model for requirements classification. The proposed model combines BERT with convolution and pooling layers. The complete matrix of token representations from the final BERT model is inputted into the CNN network [9].

Chen et al. enhanced Chinese news categorization with the LFCN model, treating it as a long text classification task. An algorithm extracts short text pairs from the lengthy input text. Subsequently, BERT is utilized for text embedding, followed by the application of a convolution-based layer to extract crucial local features [10].

Abas et al. employed a BERT-CNN for emotion detection in SemEval-2019. Their approach involves sending the BERT output to a CNN comprising a convolutional layer, pooling layer, dropout layer, dense layer, and output layer [4]. Ouni et al. developed a spam detection model for the social media domain, utilizing BERT and a topic model for the initial representation. Following this, a CNN classifier with three filter sizes is employed for classification [11].

Safaya et al. utilized BERT-CNN for Offensive Speech Identification in Social Media for SemEval-2020. They utilized the output of the last four hidden layers of all tokens as the text embedding. Subsequently, convolutions with five different sizes (768x1, 768x2, 768x3, 768x4, and 768x5) are applied to the text embedding. Finally, ReLU, Global Average Pooling, and a dense layer are employed for classification [3].

Our BERT-CNN method is part of the BERT-CNN pipeline and utilizes the last four representations of layers, similar to Safaya's work [3]. The distinguishing feature of our work compared to others is the incorporation of dilation convolution, which facilitates global feature extraction. Thus, the combination of vanilla convolution and dilation convolution allows for the extraction of both local and global features simultaneously.

In a parallel line of research, a BERT-LSTM model is introduced. Rai et al. developed a BERT-LSTM for fake news classification, incorporating an LSTM layer and an MLP after the BERT model [12]. Pandey et al. proposed a BERT-LSTM for sarcasm detection in code-mixed social media, consisting of BERT, LSTM, and a dense layer in sequence [13]. Pham-Hong and Chokshi employed the BERT-LSTM model with the Noisy Student method for multilingual offensive language identification in social media at SemEval-2020. The architecture of their model utilizes BERT for token embedding, with the outputs of text token embeddings sent to an LSTM network. The LSTM's output and the BERT embedding of the [CLS] token are concatenated and inputted to an MLP for classification [14].

Cai et al. utilized a BERT-BiLSTM model for sentiment analysis in the energy market domain, where the outputs of all tokens, including the [CLS] token from BERT, are inputted into a BiLSTM for the final text embedding [15]. Phan et al. developed a model based on BERT, CNN, BiLSTM, and GCN for aspect-level sentiment analysis. BERT-BiLSTM creates contextualized word representations, followed by GCN extracting significant features. In the final stage, CNN classifies the processed embedding [16]. Song et al. proposed a BERT-LSTM and a BERT-Attention for sentiment analysis in an aspect-oriented manner. These models extract embeddings of the [CLS] token from all layers. The BERT-LSTM inputs the [CLS] embeddings to an LSTM pooling, and the BERT-Attention attends from a learnable fixed query to embeddings [17]. This work serves as inspiration for our BERT-LSTM model.

B. Persian Social Media Analysis

Some research focuses on Persian sentiment analysis. Gasemi et al. proposed cross-lingual Persian sentiment analysis, employing a sentence-aligned supervised approach, BilBOWA, and an orthogonal-based word-aligned approach, VecMap, for cross-lingual embedding. They utilized a combination of CNN and LSTM for text classification [18]. Dashtipour et al. introduced a hybrid framework based on dependency grammar rules and neural networks. In this framework, the input text is parsed using a dependency parser. If dependency-based rules can detect the text's polarity, the framework returns it; otherwise, the CNN-based or LSTM-based model predicts the polarity [19]. In a subsequent research endeavor, Dashtipour et al. developed a model relying solely on deep neural networks, utilizing FastText for word embedding and a stacked BiLSTM for text classification in the optimal setting [20]. In another study, Dashtipour et al. investigated the accuracy of an ensemble method incorporating various classifiers [21].

Shumaly et al. proposed a method based on FastText as the word embedding and a CNN model for text classification [22]. Jafarian et al. employed a traditional BERT for aspect-based sentiment analysis [23]. Davar et al. proposed a model based on BERT and BiLSTM. Similar to our work, they compared some of the pooling heads. The key distinctions between our work and theirs lie in the research domain (political vs. shopping), code availability (our codes are open), and complexity (our models are dedicated and have more learnable parameters) [24]. Dehghani et al. developed a model based on BERT, 1D-Convolution, and BiLSTM for Persian political sentiment analysis [25].

Stance detection is less renowned than sentiment analysis and represents a recent research focus. In the context of the Persian language, few works have been undertaken. Zarharan et al. established the first Persian stance detection dataset. Alongside the dataset, they utilized a Stack LSTM and classic machine learning models such as Logistic Regression, SVM, Random Forest, and Naive Bayes [26]. Nasiri and Analoui employed Easy Data Augmentation as a data augmentation technique to enhance accuracy. This technique involves random deletion, random swap, random insertion, and synonym replacement [27]. Farhoodi et al. employed Bag of Words, TF-IDF, FastText, or BERT as the feature extractor. Their classification arsenal includes classic machine learning models like Logistic Regression, Decision Tree, SVM,

[DOI: 10.61186/itrc.16.3.9]

Random Forest, KNN, and Ada-boost, alongside deep learning models such as LSTM and BERT. They also introduced ensemble models and data augmentation based on translation and Easy Data Augmentation [28, 29].

In this study, we explore the performance of five different network architectures across two distinct text classification tasks on Persian social media. A key highlight of our research is the comparative analysis of these architectures across two diverse tasks. This provides valuable insights for researchers seeking to select appropriate models for their specific tasks in future studies. Additionally, some of the architectures we investigate, such as BERT-ClsAverageMax and BERT-WeightingCLS, are novel in the Persian context. Examining these models across both sentiment analysis (a simpler task involving single text) and stance detection (a more complex task involving text-pair) provides a nuanced understanding of which architecture performs better under different conditions.

Furthermore, while our five proposed architectures are based on previous works, they include custom-designed modifications. For instance, in BERT-ClsAverageMax, we utilize concatenation instead of simple summation to enhance the network's performance. In the case of our BERT-CNN, the combination of vanilla convolution and dilation convolution enables simultaneous extraction of both local and global features, thereby contributing to the effectiveness of our proposed architectures.

III. Proposed Method

This section delves into both conventional and innovative techniques for employing BERT in social media classification tasks, specifically focusing on sentiment analysis and stance detection. We scrutinize various strategies to discern the conditions under which they yield optimal results.

Initially, two tasks are introduced to assess the effectiveness of our methods. Following that, the model architectures of various methods will be elucidated along with their respective details. Traditional approaches employ the CLS token for classification, along with average pooling and max pooling for text classification. More recent methods, as highlighted in the previous section, involve the use of CNN, RNN, GCN, and transformers. These techniques are implemented in conjunction with different transformer layers in BERT.

In recent years, transfer learning has demonstrated remarkable results in natural language processing tasks like sequence classification, entity recognition, and question-answering. In our case, the knowledge embedded in the pre-trained BERT features is transferred to the downstream task, which, in this context, is text classification.

Numerous language models tailored for Persian have been introduced, including monolingual ones specific to Persian and multilingual ones trained in Persian alongside other languages. In this work, our classification task pertains to Persian social media; thus, we focused on Language Models that support the Persian language and are trained on social media posts. We employ XLM-T, a large language model specialized in Twitter, retrained on over 1 billion tweets from various languages until December 2022 [30]. The base model is XLM-RoBERTa large [31], which is multilingual and includes the Persian language. For simplicity, we refer to XLM-RoBERTa as BERT throughout this paper.

A. Task Descriptions

Sentiment analysis is a technique that leverages natural language processing, text analysis, and machine learning to identify and extract the emotional tone and attitude of a text. It aids in comprehending people's opinions, feelings, and emotions toward a topic, product, service, or event. Sentiment analysis finds applications in diverse domains such as social media, customer reviews, marketing, politics, and healthcare. In the context of a tweet post, the objective is to predict whether the sentiment is negative, positive, or neutral.

Stance detection is a natural language processing task that seeks to identify the attitude or opinion of a speaker or writer toward a given topic or claim. For example, given the main post, "The president states that the country is in the best condition in the last decade", and the reply post "Oh, yeah!!", the stance detection system would label the reply text as "against". Stance detection has various applications, including analyzing social media posts, detecting fake news, or summarizing debates.

In the case of a pair of sentences consisting of a main post and its reply post, the goal is to predict whether the stance of the reply post is negative, positive, or neutral concerning the main post. Various approaches exist for solving this task, and one fundamental approach, also known as pair text, involves placing two posts beside each other with the SEP token (or </s> in the case of XLM-RoBERTa) in between and providing it to the BERT model as input. In this work, we employ this approach to investigate BERT-X head performance on pair text as input.

B. Proposed Architectures

In this section, we introduce five different architectures utilized in this paper. All architectures are identical for the two tasks, except for the BERT input. In sentiment analysis, a [CLS] token is prepended to the sentence as a special token for classification. For stance detection, two texts are concatenated with a [SEP] token as a separator and a [CLS] token at the beginning. The subsequent sections elaborate on the input format for sentiment analysis.

1) BERT-MLP

The BERT-MLP head represents the most common approach for text classification. It involves using a pre-trained language model called BERT or its extensions to encode the input texts into vector representations, followed by applying a multi-layer perceptron (MLP) to classify the text labels. BERT is a powerful model capable of capturing contextual and semantic information in texts, while MLP, a simple yet effective neural network, can learn non-linear mappings from input features to output classes. In this architecture, as depicted in Fig. 1, two fully connected layers are applied to the BERT output CLS token, facilitating the

[DOI: 10.61186/itrc.16.3.9]

fusion of features extracted from BERT. After each fully connected layer, a nonlinear function RELU and a dropout are employed to prevent overfitting.

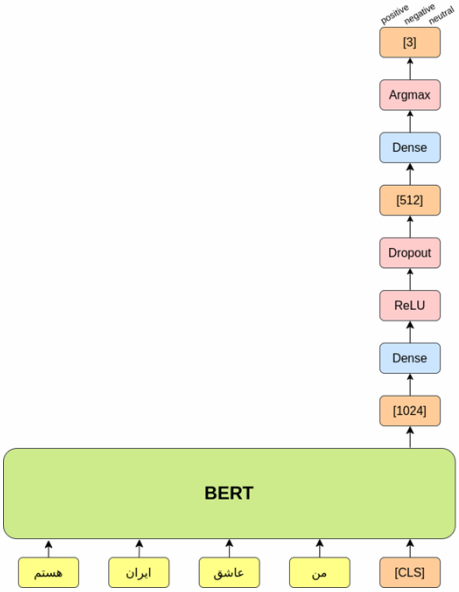

Figure 1. BERT-MLP Architecture for Sentiment Analysis

2) BERT-CNN

Various fusion approaches of BERT and CNN have led to state-of-the-art methods. Similar to the previous section, where BERT serves as vector representation, and CNN is used for extracting features and classifying texts into different categories, we introduce a novel architecture for the CNN network in this work, illustrated in Fig. 2. The representation of all tokens from the last four hidden layers is extracted and concatenated. Four CNN networks with different kernel sizes are then employed: three simple CNN kernels with distinct sizes and one dilated kernel. CNN networks with kernel sizes 2, 3, and 4 (function as 2-gram, 3-gram, and 4-gram language models) capture local word-level features. The CNN with a dilated kernel captures more global features, serving as a sentence-level feature extractor. By concatenating these four CNN networks, we leverage both local and global features together. In the final step, we employ the same MLP architecture explained in the previous section.

3) BERT-LSTM

Our BERT-LSTM architecture is presented in Fig. 3. In this design, BERT encodes texts into vector representations, and LSTM captures the sequential and temporal information in texts. The BERT-LSTM head is a common approach for text classification, utilizing BERT as the base layer and adding an LSTM layer on top. This allows the model to leverage contextual and semantic information from BERT along with long-term dependencies from LSTM.

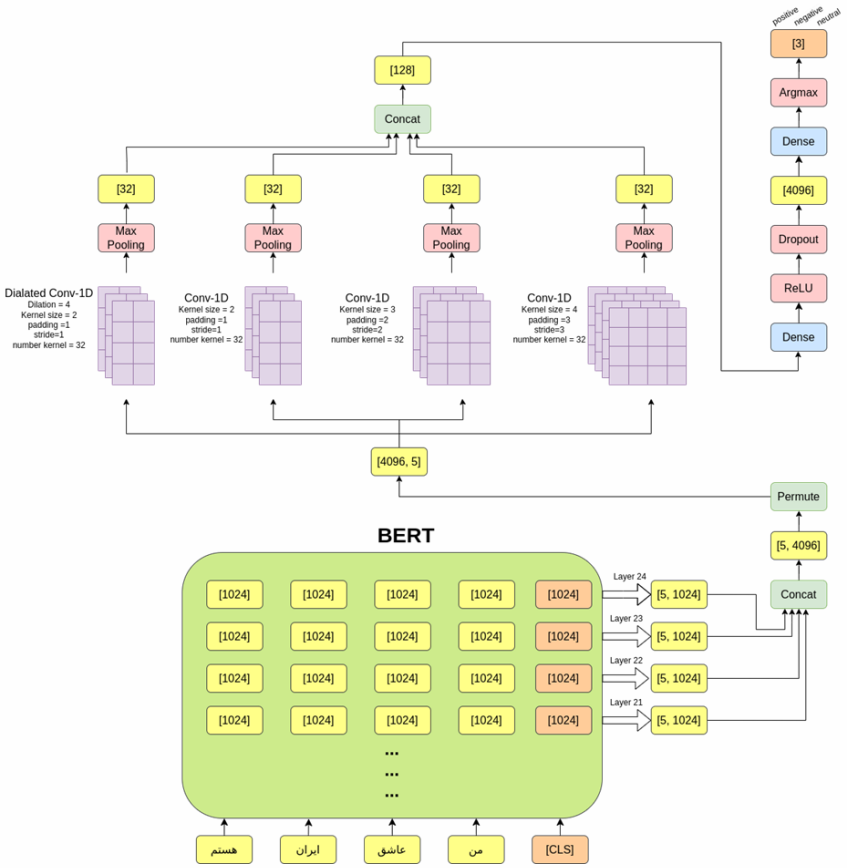

Figure 2. BERT-CNN architecture for Sentiment Analysis

[DOI: 10.61186/itrc.16.3.9]

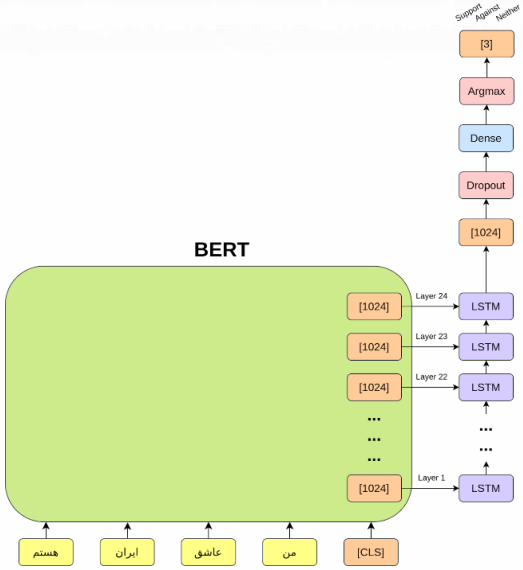

Figure 3. BERT-LSTM architecture for Sentiment Analysis

4) *BERT-WeightingCLS*

BERT-WeightingCLS utilizes the output of the CLS token from different layers of BERT, weighting and combining them to form a final representation for the text. BERT encodes texts into vector representations, and the CLS token, a special token added at the beginning of each text, captures sentence-level information. The CLS token output from different BERT layers may contain varying levels of abstraction and relevance for the text classification task. Weighting and combining these outputs may enhance the model's performance and robustness. The architecture of this model is illustrated in Fig. 4.

Different weighting schemes can be implemented for the CLS token output from different BERT layers, such as Scalar Mix, Attention, and Parameter. In this work, we use a learnable parameter for each BERT layer, multiplying the CLS token outputs by these parameters. This method allows control over the magnitude and direction of each layer's contribution to the final representation, though it may introduce overfitting or underfitting challenges.

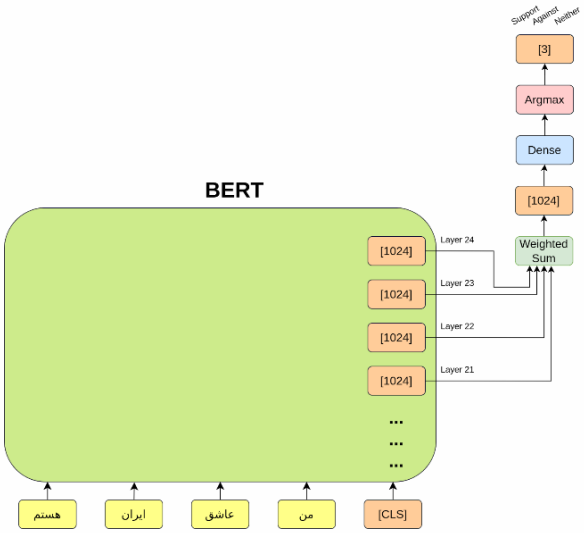

Figure 4. BERT-WeightingCLS architecture for Sentiment Analysis

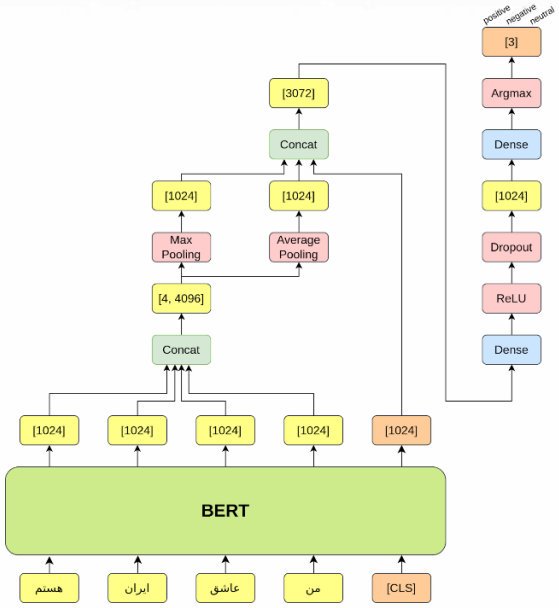

Figure 5. BERT-ClsAverageMax architecture for Sentiment Analysis

5) *BERT-ClsAverageMax*

BERT-ClsAverageMax presents another pooling layer architecture on top of BERT models for text classification, combining information from the standard [CLS] token with pooled sequence output. In our implementation shown in Fig. 5, the output of the last hidden layer is pooled in two different ways using max-pooling and average-pooling. Max-pooling captures the maximum of each feature across all tokens, emphasizing strong class-related keywords. On the other hand, average-pooling outputs the average of each feature over the sequence, attending to all tokens evenly. Finally, the features generated from the pooled output and the output for the [CLS] token are summed together.

IV. EXPRIMENTS

These methods are evaluated on the ParsiAzma competition final test dataset. The reported results in the following are based on this evaluation.

A. *Dataset*

The datasets used to train the model for this task are a combination of multiple datasets. This approach ensures that our model is trained on a larger distribution of data, preventing bias towards a specific dataset distribution. On the other hand, datasets that do not contribute to achieving better results on the test dataset are omitted, as they have different distributions.

Table 1 displays the datasets used for the training phase of sentiment analysis, while Table 2 features datasets used during the training phase of stance detection.

[DOI: 10.61186/itrc.16.3.9]

TABLE I. DATASETS FOR SENTIMENT ANALYSIS

Dataset	# Samples
Crawled Dataset (with Sabri et al. labeled [32])	2175
Emotion Test ParsiAzma (Labeled for Sentiment)	500
ParsiAzma Sample Dataset	30

TABLE II. DATASETS FOR STANCE DETECTION

Dataset	# Samples
ParsiAzma Stance dataset	4063
Crawled Dataset	1000

B. Evaluation Metrics

Precision, recall, and F1 scores are reported for each task and model. Given that both tasks here are multi-class, Macro average and Micro average are two methods commonly used to aggregate class results. In the former, precision, recall, and F1 scores are first calculated for each class, and then for each metric, the average of all classes' results is considered the final score. But, in the latter, all data points have an equal effect on the final metrics, resulting in precision, recall, and F1 scores equal to the accuracy. It is worth mentioning that, in the Macro average, if precision and recall are unbalanced for some classes, the final F1 score can be lower than both the final precision and recall scores.

C. Sentiment Analysis

Table 3 presents the results of all models for the sentiment analysis task. In terms of evaluation using F1 Macro and F1 Micro metrics, it can be generally concluded that the BERT-LSTM model exhibits superior performance compared to other models. All models achieved a higher macro F1 score than BERT-MLP as the standard BERT architecture for sentence classification. The models tend to learn better from the classes that have more samples, resulting in higher Micro scores than Macro scores. This is because the Micro score gives more weight to the large-sample classes, while the Macro score treats all classes equally.

D. Stance Detection Results

Table 4 presents the results of stance detection. It is evident that BERT-WeightingCLS has achieved the highest scores across all methods, while BERT-LSTM stands in the second rank. Similarly and based on the mentioned reason, the Micro scores are higher than the Macro scores in this task.

Figure 6 illustrates all F1 scores of all models for the two tasks. The figure demonstrates the substantial superiority of BERT-LSTM and BERT-WeightingCLS. It shows that the last layers have important and different data, and the models can benefit from these, but BERT-MLP and BERT-ClsAverageMax cannot. BERT-CNN is a complicated network that uses all tokens embedding from the last layers. Thus, this model is not suitable for these tasks with low data.

It is important to note that performing sentiment analysis is less challenging than stance detection. When dealing with straightforward tasks, the distinction between a complex network and a simpler one is negligible. Thus, nearly all models achieve comparable outcomes when it comes to sentiment analysis, which is considered an easier task. However, a more noticeable difference is observed in stance detection, which is a more complex task. In such instances, it is advantageous not only to utilize the CLS token from all layers but also to avoid adding unnecessary complexity to the network. Therefore, employing WeightingCLS can effectively meet both requirements.

E. Number of Parameters

In this section, we examine the number of parameters for each architecture. Table 5 shows the total parameters of the model and the number of parameters of the head only (the BERT parameters are not included in the Head Parameters column).

As shown in this table, the head parameters do not make a significant difference in the total number of parameters (at most 0.5 percent of the total parameters), so the training and inference phases do not change in terms of time and computational costs. Moreover, having more parameters does not necessarily lead to better results, and it depends on the task and how the BERT features are combined. BERT-LSTM and BERT-WeightingCLS, which achieve better results in both tasks, almost have fewer parameters compared to other architectures.

V. CONCLUSION

In this study, we explore the utilization of BERT-X, which combines BERT as a text feature extractor and an auxiliary network as a classification head, for Persian social media analysis. We propose five variants of BERT-X, namely BERT-MLP, BERT-CNN, BERT-LSTM, BERT-WeightingCLS, and BERT-ClsAverageMax, based on existing architectures, and evaluate them on two tasks: sentiment analysis and stance detection. Our results demonstrate that BERT-X with a complex network can outperform BERT-X with a simple MLP. Moreover, we found that BERT-LSTM and BERT-WeightingCLS achieved the highest performance scores on the two tasks. For sentiment analysis, BERT-LSTM obtains a macro F1 score of 64.56% and a micro F1 score of 73.00%. For stance detection, BERT-WeightingCLS achieves a macro F1 score of 66.30% and a micro F1 score of 71.58%.

For future work, we suggest the following directions:

- Developing more variants of BERT-X and comparing them with our proposed architectures
- Conducting more experiments on other Persian social analysis tasks, such as emotion recognition
- Extending our work to other languages and presenting cross-lingual experiments

Analyzing the reasons behind the different performance of BERT-X variants

[DOI: 10.61186/itrc.16.3.9]

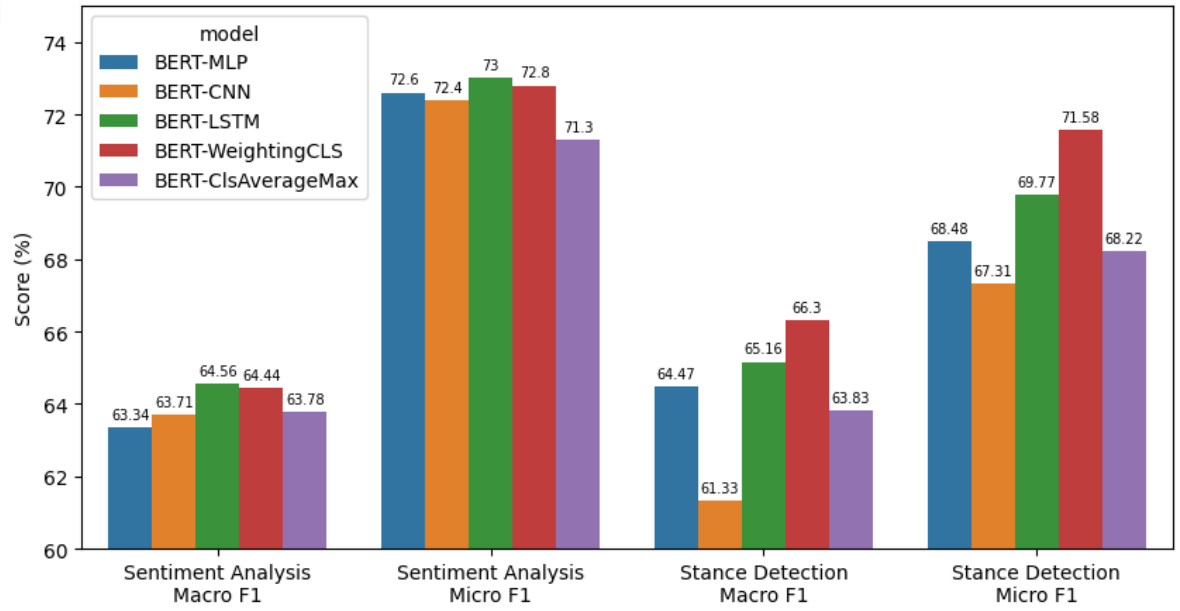

Figure 6. Comparison of all models for two tasks and two metrics

TABLE III. RESULTS OF BERT-X MODELS FOR SENTIMENT ANALYSIS

	Macro			Micro		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
BERT-MLP	64.19	68.12	63.34	72.60	72.60	72.60
BERT-CNN	**65.27**	69.39	63.71	72.40	72.40	72.40
BERT-LSTM	65.23	69.83	**64.56**	**73.00**	**73.00**	**73.00**
BERT-WeightingCLS	65.22	70.39	64.44	72.80	72.80	72.80
BERT-ClsAverageMax	65.16	**71.08**	63.78	71.30	71.30	71.30

TABLE IV. RESULTS OF BERT-X MODELS FOR STANCE DETECTION

	Macro			Micro		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
BERT-MLP	64.68	65.27	64.47	68.48	68.48	68.48
BERT-CNN	61.77	61.30	61.33	67.31	67.31	67.31
BERT-LSTM	64.74	65.67	65.16	69.77	69.77	69.77
BERT-WeightingCLS	**67.16**	**65.72**	**66.30**	**71.58**	**71.58**	**71.58**
BERT-ClsAverageMax	63.32	64.58	63.83	68.22	68.22	68.22

TABLE V. NUMBER OF PARAMETERS

Model Architecture	**Head Parameters**	**Total Parameters**
BERT-MLP	1,052,675	560,943,107
BERT-CNN	1,577,091	561,467,523
BERT-LSTM	1,313,539	561,203,971
BERT-WeightingCLS	3,099	559,893,531
BERT-ClsAverageMax	3,149,827	563,040,259

[DOI: 10.61186/itrc.16.3.9]

REFERENCES

[1] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[2] Dong, Junchao, Feijuan He, Yunchuan Guo, and Huibing Zhang. "A commodity review sentiment analysis based on BERT-CNN model." In *2020 5th International conference on computer and communication systems (ICCCS)*, pp. 143-147. IEEE, 2020.

[3] Safaya, Ali, Moutasem Abdullatif, and Deniz Yuret. "Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media." *arXiv preprint arXiv:2007.13184* (2020).

[4] Abas, Ahmed R., Ibrahim Elhenawy, Mahinda Zidan, and Mahmoud Othman. "BERT-CNN: A Deep Learning Model for Detecting Emotions from Text." *Computers, Materials & Continua* 71, no. 2 (2022).

[5] Lehecka, Jan, Jan Svec, Pavel Ircing and Lubos Smídl. "Adjusting BERT's Pooling Layer for Large-Scale Multi-Label Text Classification." Workshop on Time-Delay Systems (2020).

[6] Zheng, Shaomin, and Meng Yang. "A new method of improving bert for text classification." In Intelligence Science and Big Data Engineering. Big Data and Machine Learning: 9th International Conference, IScIDE 2019, Nanjing, China, October 17–20, 2019, Proceedings, Part II 9, pp. 442-452. Springer International Publishing, 2019.

[7] Wan, Chang-Xuan, and Bo Li. "Financial causal sentence recognition based on BERT-CNN text classification." *The Journal of Supercomputing* (2022): 1-25.

[8] Jia, Keliang. "Sentiment classification of microblog: A framework based on BERT and CNN with attention mechanism." Computers and Electrical Engineering 101 (2022): 108032.

[9] Kaur, Kamaljit, and Parminder Kaur. "BERT-CNN: Improving BERT for Requirements Classification using CNN." *Procedia Computer Science* 218 (2023): 2604-2611.

[10] Chen, Xinying, Peimin Cong, and Shuo Lv. "A long-text classification method of Chinese news based on BERT and CNN." *IEEE Access* 10 (2022): 34046-34057.

[11] Ouni, Sarra, Fethi Fkih, and Mohamed Nazih Omri. "BERT- and CNN-based TOBEAT approach for unwelcome tweets detection." Social Network Analysis and Mining 12, no. 1 (2022): 144.

[12] Rai, Nishant, Deepika Kumar, Naman Kaushik, Chandan Raj, and Ahad Ali. "Fake News Classification using transformer based enhanced LSTM and BERT." *International Journal of Cognitive Computing in Engineering* 3 (2022): 98-105.

[13] Pandey, Rajnish, and Jyoti Prakash Singh. "BERT-LSTM model for sarcasm detection in code-mixed social media post." Journal of Intelligent Information Systems 60, no. 1 (2023): 235-254.

[14] Pham-Hong, Bao-Tran, and Setu Chokshi. "PGSG at SemEval-2020 task 12: BERT-LSTM with tweets' pretrained model and noisy student training method." In Proceedings of the fourteenth workshop on semantic evaluation, pp. 2111-2116. 2020.

[15] Cai, Ren, Bin Qin, Yangken Chen, Liang Zhang, Ruijiang Yang, Shiwei Chen, and Wei Wang. "Sentiment analysis about investors and consumers in energy market based on BERT-BiLSTM." IEEE access 8 (2020): 171408-171415.

[16] Phan, Huyen Trang, Ngoc Thanh Nguyen, and Dosam Hwang. "Aspect-level sentiment analysis using CNN over BERT-GCN." *IEEE Access* 10 (2022): 110402-110409.

[17] Song, Youwei, Jiahai Wang, Zhiwei Liang, Zhiyue Liu, and Tao Jiang. "Utilizing BERT intermediate layers for aspect based sentiment analysis and natural language inference." arXiv preprint arXiv:2002.04815 (2020).

[18] Ghasemi, Rouzbeh, Seyed Arad Ashrafi Asli, and Saeedeh Momtazi. "Deep Persian sentiment analysis: Cross-lingual training for low-resource languages." *Journal of Information Science* 48, no. 4 (2022): 449-462.

[19] Dashtipour, Kia, Mandar Gogate, Jingpeng Li, Fengling Jiang, Bin Kong, and Amir Hussain. "A hybrid Persian sentiment analysis framework: Integrating dependency grammar based rules and deep neural networks." Neurocomputing 380 (2020): 1-10.

[20] Dashtipour, Kia, Mandar Gogate, Ahsan Adeel, Hadi Larijani, and Amir Hussain. "Sentiment analysis of Persian movie reviews using deep learning." *Entropy* 23, no. 5 (2021): 596.

[21] Dashtipour, Kia, Cosimo Ieracitano, Francesco Carlo Morabito, Ali Raza, and Amir Hussain. "An ensemble based classification approach for Persian sentiment analysis." Progresses in Artificial Intelligence and Neural Systems (2021): 207-215.

[22] Shumaly, Sajjad, Mohsen Yazdinejad, and Yanhui Guo. "Persian sentiment analysis of an online store independent of pre-processing using convolutional neural network with fastText embeddings." PeerJ Computer Science 7 (2021): e422.

[23] Jafarian, Hamoon, Amir Hossein Taghavi, Alireza Javaheri, and Reza Rawassizadeh. "Exploiting BERT to improve aspect-based sentiment analysis performance on Persian language." In 2021 7th International Conference on Web Research (ICWR), pp. 5-8. IEEE, 2021.

[24] Davar, Omid, Gholamreza Dar, and Fahimeh Ghasemian. "DeepSentiParsBERT: A Deep Learning Model for Persian Sentiment Analysis Using ParsBERT." In *2023 28th International Computer Conference, Computer Society of Iran (CSICC)*, pp. 1-5. IEEE, 2023.

[25] Dehghani, Mohammad, and Zahra Yazdanparast. "Sentiment Analysis of Persian Political Tweets Using ParsBERT Embedding Model with Convolutional Neural Network." In 2023 9th International Conference on Web Research (ICWR), pp. 20-25. IEEE, 2023.

[26] Zarharan, Majid, Samane Ahangar, Fateme Sadat Rezvaninejad, Mahdi Lotfi Bidhendi, Mohammad Taher Pilevar, Behrouz Minaei, and Sauleh Eetemadi. "Persian Stance Classification Data Set." In TTO. 2019.

[27] Nasiri, Homa, and Morteza Analoui. "Persian stance detection with transfer learning and data augmentation." In 2022 27th International Computer Conference, Computer Society of Iran (CSICC), pp. 1-5. IEEE, 2022.

[28] Farhoodi, Mojgan, Abbas Toloie Eshlaghy, and M. R. Motadel. "A Proposed Model for Persian Stance Detection on Social Media." *International Journal of Engineering* 36, no. 6 (2023): 1048-1059.

[29] Farhoodi, Mojgan, Abbas Toloie Eshlaghy, and Mohamadreza Motadel. "The Effect of Data Augmentation Techniques on Persian Stance Detection." International Journal of Information and Communication Technology Research 15, no. 1 (2023): 63-71.

[30] Barbieri, F., Anke, L. E., & Camacho-Collados, J. (2022). XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. 2022 Language Resources and Evaluation Conference, LREC 2022, 258–266.

[31] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[32] Sabri, Nazanin, Reyhane Akhavan, and Behnam Bahrak. "Emopars: A collection of 30k emotion-annotated persian social media texts." In Proceedings of the Student Research Workshop Associated with RANLP 2021, pp. 167-173. 2021.

Mohamad Sobhi received his M.Sc. degree in Software Engineering from Sharif University of Technology in 2021. He is currently a Ph.D. candidate at Amirkabir Univesity of technology. His current interest is Natural Language Understanding, Natural Language generation, Reinforcement Learning and Complex Networks. He is currently

[DOI: 10.61186/itrc.16.3.9]

working on Developing dialogue system with Large Language Models(LLM) and Reinforcement Learning.

Alireza Mazochi received his B.Sc. degree in Computer Engineering in 2021 and his M.Sc. degree in Artificial Intelligence in 2024, both from Amirkabir University of Technology. His research interests include Natural Language Processing (NLP), Large Language Models (LLM), Chatbots, Graph Neural Networks, and Information Retrieval.

Hossein Zeinali is an assistant professor in the Artificial Intelligence group of the Computer Engineering Department at Amirkabir University of Technology. He is the director of the Speech and Language Technologies (SLT) lab in the department. He received his B.Sc. degree in Computer Engineering from Shiraz University, Iran, in 2010, and his M.Sc. and Ph.D. degrees in Artificial Intelligence from Sharif University of Technology, Tehran, Iran, in 2012 and 2017, respectively. He was a visiting student and a postdoctoral researcher at the Speech Group of Brno University of Technology, Czech Republic. His research interests include speech and speaker recognition, speech-to-text, dialog systems, chatbots, and large language models.

[DOI: 10.61186/itrc.16.3.9]