Journal of Innovations in Computer Science and Engineering
Shahid Beheshti Unversity

June 2024, Special Issue on AI 4 All- 1

Intermediate Fine-Tuning for Robust Persian Emotion Detection in Text

Morteza Mahdavi Mortazavi✉,, Code ORCID: 0009-0007-1915-7787
Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran. s.mahdavimortazavi@sbu.ac.ir
Mehrnoush Shamsfard, Code ORCID: 0000-0002-7027-7529
Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran. m-shams@sbu.ac.ir

Abstract

Emotion recognition in text is a growing area in Natural Language Processing (NLP), essential for improving human-computer interactions by allowing systems to interpret emotional expressions. While much progress has been made in English, Persian emotion recognition has seen limited devel- opment due to resource constraints and linguistic challenges. In this study, we address these gaps by leveraging two key Persian datasets, ArmanEmo and ShortEmo, to train an efficient emotion recognition model. Using FaBERT, a BERT-based model optimized for Persian, we employ interme- diate fine-tuning on a large collection of informal and formal Persian texts to enhance the model's adaptability to colloquial language. This step significantly improves comprehension of Persian text variations, as reflected in reduced perplexity scores. Our final evaluations, incorporating accuracy, precision, recall, and F1 score metrics, demonstrate that this fine-tuned FaBERT model achieves strong performance in emotion recognition, providing a promising approach for NLP in low-resource languages.

Keywords: Emotion Recognition, Persian Text Processing, Intermediate Fine-Tuning, BERT-based Models, NLP in Low-resource Languages.

1 Introduction

1.1 Background

Emotion recognition in text is a pivotal task in Natural Language Processing (NLP) that involves iden- tifying and interpreting human emotions expressed through language. Emotions are psychological states triggered by internal or external stimuli and are often associated with thoughts, feelings, and behaviors. They can be categorized into discrete classes such as anger, sadness, happiness, fear, hate, surprise, and neutral (other). Recognizing these emotions enables applications such as sentiment analysis, customer feedback analysis, mental health monitoring, and personalized human-computer interactions.

While emotion recognition in English has seen considerable progress due to the abundance of annotated datasets and resources, emotion recognition for Persian text lags behind due to the scarcity of high- quality annotated datasets, the structural complexity of the Persian language, and cultural nuances that influence emotional expression [1, 17].

Recent efforts have introduced resources like the ArmanEmo and ShortPersianEmo datasets to address these gaps. ArmanEmo provides over 7,000 Persian sentences labeled across seven emotional categories: angry, sad, hate, surprise, fear, happy, and neutral (other) [1]. Similarly, ShortPersianEmo offers 5,472 concise Persian text samples annotated for five main emotions: angry, sad, fear, happy, and neutral (other) [17]. These datasets represent a mix of formal and informal text, including social media posts and user reviews, making them valuable resources for advancing Persian NLP.

1.2 Our Work

In this study, we propose a novel approach for Persian text-based emotion recognition by leveraging FaBERT, a BERT-based model optimized for Persian. Our contributions are as follows:

- We conduct intermediate fine-tuning of FaBERT on a large corpus of informal Persian text, includ- ing user-generated content from social media and reviews, to enhance its adaptability to colloquial language.
- We evaluate the model's performance on two benchmark datasets, ArmanEmo and ShortPer- sianEmo, using metrics such as accuracy, precision, recall, and F1 score, and demonstrate that our approach outperforms existing methods.
- We address key challenges, such as class imbalance and underrepresentation, through data aug- mentation and careful preprocessing techniques to ensure robust training.

This paper discusses related work, describes the datasets and fine-tuning methodology, and provides a detailed

Submit Date: 2024-10-30
Accept Date: 2025-05-05
✉ Corresponding author

analysis of the experimental results. It highlights the impact of intermediate fine-tuning and preprocessing strategies, particularly in enhancing the detection of emotions in informal Persian texts. The findings demonstrate significant advancements in Persian emotion recognition, with notable improvements on the challenging and nuanced ArmanEmo dataset.

2 Related Work

2.1 Early Approaches

Initial methods for emotion recognition in text relied on rule-based and lexicon-based approaches, uti- lizing predefined linguistic patterns and emotion lexicons to infer emotions [6]. While these approaches were interpretable and easy to implement, they often lacked sensitivity to context and struggled with ambiguity, leading to misclassifications and limited scalability across diverse domains [2].

2.2 Transition to Machine Learning Techniques

To overcome the limitations of rule-based methods, machine learning models such as Naive Bayes, Sup- port Vector Machines (SVMs), and Decision Trees were introduced. These models leveraged engineered features like n-grams and part-of-speech tags to improve classification accuracy [6, 11]. However, they required extensive manual feature engineering and were unable to capture deep semantic relationships, limiting their performance in complex emotion recognition tasks.

2.3 Advancements with Deep Learning and Transformers

The advent of deep learning brought significant advancements in emotion recognition. Models like Long Short-Term Memory networks (LSTMs) and Convolutional Neural Networks (CNNs) improved upon traditional methods by capturing sequential dependencies in text [11, **?**]. Despite these improvements, challenges such as long-term dependency modeling and computational complexity persisted.

Transformers, particularly BERT, revolutionized the field by providing richer contextual representa- tions through attention mechanisms [19]. Fine-tuned models like ParsBERT, pretrained on Persian text, achieved state-of-the-art performance on datasets such as ArmanEmo and EmoPars, demonstrating their effectiveness for Persian emotion detection [1, 3, 9].

2.4 Non-English and Low-Resource Languages

Emotion recognition in low-resource languages, such as Persian, has faced challenges due to limited annotated datasets and pretrained models. Recent contributions like ArmanEmo and EmoPars have provided annotated datasets specifically for Persian, addressing this gap [1, 3]. ParsBERT, when fine- tuned on these datasets, demonstrated strong performance, showcasing the potential of transfer learning for low-resource settings [9].

2.5 Challenges and Gaps in Emotion Recognition

Despite advancements, several challenges persist:

- **Overlapping Emotions:** Emotions such as sadness and anger often coexist, complicating single- label classification [1].
- **Context Sensitivity:** Emotional interpretation varies significantly depending on context [3].
- **Subjectivity:** Different individuals may perceive the same text differently, introducing ambiguity [2].
- **Limited Data:** Annotated datasets for Persian emotion recognition remain sparse [1].
- **Cultural Variability:** Emotional expressions differ across cultures, complicating cross-lingual applications [19].
- **Figurative Language:** Idioms, metaphors, and sarcasm make it difficult to extract literal emo- tional content [17].

3 Methodology

In this section, we outline the structured methodology employed for emotion recognition in Persian text. This includes the datasets utilized, preprocessing steps, model architecture, training and fine-tuning procedures, and evaluation metrics.

3.1 Datasets

We utilized two primary Persian datasets, ArmanEmo and ShortEmo, for training and evaluating our model. To enhance adaptability to informal Persian text, we performed intermediate fine-tuning on a diverse collection of datasets. These datasets are described below:

3.1.1 *Intermediate Fine-Tuning Datasets*

To address the gap between formal and informal Persian text, intermediate fine-tuning was conducted on a curated collection of datasets. This step exposed the model to diverse language styles, emphasizing informal expressions. A total of 6 million Persian sentences were used, with 500,000 samples gathered from datasets such as Snapp Tweets, DigikalaMag, and Instagram Sentiment Analysis [3, 1, 21].

The remaining 5.5 million sentences were derived from a large-scale colloquial Persian corpus, providing a rich mixture of formal and informal content. This approach ensured robustness in handling diverse linguistic patterns and topics [19].

3.2 Focus on Informal Text and Colloquialism

The intermediate fine-tuning datasets predominantly consisted of user-generated content from social media, reviews, and informal conversations. Key datasets include:

- **Snapp Tweets:** 22,601 tweets reflecting user opinions and informal language.
- **Instagram Comments:** 111,733 comments providing insights into colloquial expressions and emojis.
- **DigikalaMag Reviews:** 8,515 reviews with informal user-generated content.
- **Large-Scale Colloquial Persian Corpus:** 5.5 million sentences emphasizing informal expres- sions.

The curated datasets were instrumental in equipping the model with the ability to handle unstructured and conversational Persian text. The inclusion of diverse sources ensured a balanced representation of formal and informal language styles [3, 21].

3.3 Preprocessing

Our preprocessing pipeline closely follows the methodology used in Mirzaee et al. (EmoRecBiGru) with modifications to better suit our data and specific research goals. These adjustments aimed to preserve valuable features in the text that may contribute to the emotional context, as detailed below:

- **Remove Arabic Diacritics:** Diacritics were removed to standardize Persian text.
- **Remove Non-Persian Characters:** Non-Persian characters, including some Arabic letters, were filtered out.
- **Correct Repeated Characters:** Excessive character repetition was minimized while preserving emphasis.
- **Remove English Characters:** English letters were removed, but hashtags were retained.
- **HTML Cleaning:** HTML tags were stripped, retaining valuable content like URLs.
- **Text Normalization:** Hazm was used for standardizing text, improving token consistency.
- **Emoji Handling:** Emojis were preserved for their emotional value.

3.4 Model Architecture

Our model architecture is based on FaBERT, a BERT-based language model fine-tuned specifically for Persian NLP tasks. FaBERT was chosen due to its efficiency and effectiveness, providing competi- tive performance with fewer parameters than XLM-Roberta-Large. Despite having significantly fewer parameters, FaBERT's performance in emotion recognition has been shown to match that of XLM- Roberta-Large, making it a highly efficient choice.

3.5 Intermediate Fine-Tuning

To improve FaBERT's performance on informal Persian text, we employed an intermediate fine-tuning step, continuing the model's pretraining with a 25% masking rate. This decision was informed by experiments evaluating the effects of different masking rates (15%, 20%, 25%, and 30%) on training dynamics and downstream performance, as well as insights from prior research.

Intermediate fine-tuning, also known as continued pretraining, is a pivotal strategy for enhancing lan- guage models' performance on domain-specific and informal text. This process involves further training a pretrained model on domain-relevant data, enabling it to better capture the nuances of specialized and colloquial language.

3.5.1 *Rationale for Intermediate Fine-Tuning*

- **Domain Adaptation:** Pretrained language models are typically trained on large, diverse corpora that may not adequately represent specific domains or informal language. Intermediate fine-tuning on domain-specific and informal text allows the model to adapt to the unique linguistic patterns and vocabulary prevalent in such data, thereby improving its performance on related tasks [19].
- **Handling Linguistic Variations:** Informal language often includes colloquialisms, slang, and non-standard grammar. By exposing the model to these variations during intermediate fine-tuning, it becomes more adept at understanding and generating informal text [20].
- **Improved Performance Metrics:** Studies have shown that intermediate fine-tuning can lead to significant improvements in performance metrics such as perplexity, which measures the model's uncertainty in predicting the next word. Lower perplexity indicates a better understanding of the language, which is particularly beneficial when dealing with domain-specific and informal text [21].

3.5.2 *Supporting Evidence*

A study by Gururangan et al. (2020) demonstrated that continued pretraining on domain-specific corpora led to improved performance on downstream tasks within that domain, highlighting the effectiveness of this approach [19]. Similarly, Chang and Lu (2021) explored the impact of intermediate-task fine-tuning and found that it enhances the model's ability to handle domain-specific and informal language tasks [20]. Furthermore, Alghanmi et al. (2022) proposed a self-supervised intermediate fine-tuning strategy for biomedical language models, focusing on interpreting patient case descriptions. Their approach involved fine-tuning models on the task of predicting masked medical concepts from PubMed abstracts,
leading to substantial performance improvements on biomedical NLP tasks [21].

3.5.3 *Masking Rate Analysis and Prior Research*

During the intermediate fine-tuning phase, we evaluated the impact of varying masking rates (15%, 20%, 25%, and 30%) using consistent configurations: 40,000 training steps, a batch size of 32, and a sequence length of 512 tokens.

Our findings align with observations from the paper *"Should You Mask 15% in Masked Language Modeling?"*, which highlights the potential benefits of higher masking rates. That study demonstrated that larger models, such as BERT-large, can achieve improved performance with masking rates as high as 40%. However, it also cautions against excessively high masking rates, as they may hinder convergence and representation learning in smaller models.

For FaBERT, which is a smaller model compared to competitors like XLM-Roberta-large or BERT- large, our experiments revealed that masking rates above 30% introduced training instability. Specifically:

- **30% Masking Rate:** This configuration resulted in a perplexity score of 11, significantly higher than the scores achieved with lower masking rates. The model struggled to converge effectively, likely due to excessive removal of context, which overwhelmed its capacity to reconstruct masked tokens.
- **15% and 20% Masking Rates:** These rates achieved perplexity scores of 7.1 and 7.3, respec- tively, slightly lower than the 7.5 perplexity observed with the 25% masking rate. However, the downstream performance of these configurations was marginally inferior, suggesting that lower masking rates did not introduce sufficient variability to maximize the model's learning potential.
- **25% Masking Rate:** This rate achieved a balance between contextual richness and learning difficulty, yielding a perplexity score of 7.5. Despite the slightly higher perplexity, models fine-tuned with this masking rate consistently achieved superior downstream task performance, outperforming other configurations by approximately 0.2% in F1 score.

3.6 Final Training and Fine-Tuning

After intermediate fine-tuning, we proceeded to train FaBERT on ArmanEmo and ShortEmo for emotion classification.

3.6.1 *Training Configuration*

The configuration for this final fine-tuning was as follows: **Epochs:** 7, **Batch Size:** 32, **Max Sequence Length:** 256, **Learning Rate:** 1e-5.

3.7 Evaluation Metrics

Evaluation metrics included: **Accuracy**, **Precision**, **Recall**, **F1 Score**.

Due to dataset imbalance, we used macro-averaging, computing each metric independently for each class and then taking the average to ensure balanced evaluation across all emotion categories.

Why Choose Macro Averaging for Imbalanced Datasets: In imbalanced datasets, where cer- tain classes have significantly fewer instances than others, the choice of averaging method impacts the evaluation:

- **Equal Importance to All Classes:** Macro averaging ensures that each class contributes equally to the final metric, preventing dominant classes from overshadowing minority ones [18].
- **Highlighting Performance on Minority Classes:** By treating all classes equally, macro aver- aging can reveal if the model performs poorly on less frequent classes, which might be masked in weighted or micro averaging [18].
- **Avoiding Bias Towards Majority Classes:** Weighted and micro averaging can lead to high overall scores even if the model fails to predict minority classes accurately, due to their emphasis on majority classes [18].

Supporting Evidence: Research indicates that macro-averaged metrics are more informative in sce- narios with class imbalance:

> "In the context of an imbalanced dataset where equal importance is attributed to all classes, opting for the macro average stands as a sound choice since it treats each class with equal significance." [18]
>
> "Macro-average F1 score is computed by taking the arithmetic mean (aka unweighted mean) of all the per-class F1 scores. This method treats all classes equally regardless of their support values." [18]

4 Results And Discussions

The performance metrics of FaBERT across various configurations and benchmarks are summarized below. To facilitate clarity, results for the ArmanEmo and ShortEmo datasets are presented in separate tables. These comparisons include results from the ArmanEmo and ShortEmo papers as benchmarks, alongside the performance of FaBERT.

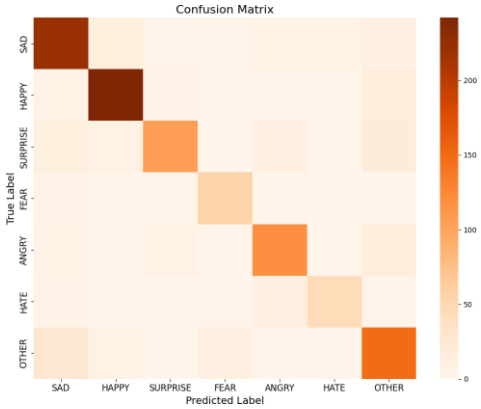
4.1 ArmanEmo Results

Table 1 provides a detailed comparison of various models on the ArmanEmo dataset. FaBERT shows substantial improvements over previous models, particularly with intermediate fine-tuning.

Table 1: Performance Comparison of BERT-based Models

Model Name	Precision (Macro)	Recall (Macro)	F1 (Macro)
ParsBERT	67.10	65.56	65.74
XLM-Roberta-base	72.26	68.43	69.21
XLM-Roberta-large	75.91	75.84	75.39
XLM-EMO-t	70.05	68.08	68.57
FaBERT-Base	75.67	75.67	75.67
FaBERT-Intermediate FT	79.64	77.96	78.47

Figure 1: Confusion Matrix of the Model's Predictions. The matrix illustrates the distribution of pre- dicted versus actual emotion labels, for ArmanEmo Test



4.2 ShortEmo Results

Table 2 presents the performance of various models on the ShortEmo dataset. For models that were not evaluated in the ShortEmo paper, placeholder values have been added.

Table 2: Performance Comparison of Models

Model Name	Macro-F1	Accuracy
ParsBERT	71%	73%
FaBERT-Base	72.2%	73.9%
FaBERT-Intermediate FT	73.0%	75.1%

FaBERT-Base achieves an F1 score of 72.2%, marginally surpassing ParsBERT's 71%. With interme- diate fine-tuning, FaBERT achieves a modest improvement, reaching an F1 score of 73.0%. The gains in ShortEmo are less pronounced compared to ArmanEmo, reflecting the simpler and more formal nature of the ShortEmo dataset.

4.3 Analysis

Across both datasets, FaBERT demonstrates strong performance, particularly on the more complex Ar- manEmo dataset, where intermediate fine-tuning on informal data provides significant benefits. The results confirm the importance of tailoring fine-tuning strategies to the linguistic and stylistic character- istics of the target dataset. FaBERT consistently outperforms most prior models, with the exception of XLM-Roberta-large on ShortEmo, where the differences are negligible.

These findings validate FaBERT as a robust and efficient solution for Persian emotion recognition tasks, advancing the state-of-the-art in this domain.

5 Conclusion

The study addresses the critical challenge of emotion recognition in Persian text, focusing on the com- plexities of informal language. Informal text, characterized by slang, idioms, non-standard grammar, and colloquial expressions, presents significant obstacles for conventional fine-tuning methods, which are typically optimized for more formal text. These challenges are further compounded by the limited availability of annotated Persian datasets and the linguistic intricacies of the language.

To overcome these limitations, the research introduces an intermediate fine-tuning approach for FaBERT, a Persian-optimized BERT-based model. This method involves pretraining FaBERT on a curated col- lection of predominantly informal Persian texts, including user-generated content from social media, reviews, and conversational sources. By doing so, the model becomes adept at handling the unstruc- tured and conversational nature of informal Persian language. The intermediate fine-tuning step reduced perplexity and significantly improved downstream performance.

Experimental results demonstrated substantial improvements in macro-F1 scores for FaBERT, par- ticularly on datasets like ArmanEmo, which involve complex and nuanced informal text. The refined FaBERT model outperformed traditional fine-tuning strategies and even rivaled larger models like XLM- RoBERTa in effectiveness, despite having fewer parameters.

In summary, this study underscores the importance of adapting language models to informal text through intermediate fine-tuning, showcasing a practical and effective solution for emotion recognition in low-resource languages like Persian. The approach bridges a critical gap, achieving significant perfor- mance gains in scenarios where standard fine-tuning falls short.

References

[1] E. Sadeghi, S. Hosseini, M. Yazdani, and S. H. Akhavan, "ArmanEmo: A Human-Labeled Emotion Dataset for Persian Language," *arXiv preprint*, arXiv:2207.11808, 2022.

[2] M. Keshavarz and H. Faili, ”Automatic Persian Text Emotion Detection,” *Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media (PEOPLES)*, 2018.
[3] F. Tavakoli, M. Akbari, and M. Zarei, ”EmoPars: A Dataset for Emotion Recognition in Persian,” *arXiv preprint*, arXiv:2111.07815, 2021.
[4] A. Jafari, A. Tajik, and Z. Nemat, ”Emotion Analysis of Tweets Banning Education in Afghanistan,” *arXiv preprint*, arXiv:2204.07556, 2022.
[5] S. Salehi, A. Mehranfar, and M. Mansouri, ”Deep Emotion Detection and Sentiment Analysis of Persian Literary Text,” *Digital Scholarship in the Humanities*, vol. 37, no. 2, pp. 245–265, 2022.
[6] M. Poria, E. Cambria, A. Hussain, and G. Huang, ”Rule-Based Systems for Emotion Detection,” *Cognitive Computation*, vol. 7, no. 3, pp. 243–259, 2015.
[7] T. Wettig, C. Snyder, and H. Li, ”Should You Mask 15% in Masked Language Modeling?,” *arXiv preprint*, arXiv:2202.08005, 2022.
[8] H. Mirzaee et al., ”ArmanEmo: A Persian Dataset for Text-based Emotion Detection,” *arXiv preprint*, arXiv:2207.11808, 2022.
[9] A. Abaskohi et al., ”Persian Emotion Detection using ParsBERT and Imbalanced Data Handling Approaches,” *arXiv preprint*, arXiv:2211.08029, 2022.
[10] S. Sadeghi et al., ”Automatic Persian Text Emotion Detection using Cognitive Linguistic and Deep Learning,” *Shahrood University Journals*, 2021.
[11] Rasouli et al., ”Investigating Shallow and Deep Learning Techniques for Emotion Classification in Short Persian Texts,” *JAD Shahrood University*, 2021.
[12] S. Rahmani Zardak et al., ”Persian Text Sentiment Analysis Based on BERT and Neural Networks,” *SpringerLink*, 2023.
[13] A. Yazdani et al., ”Emotion Recognition in Persian Speech Using Deep Neural Networks,” *arXiv preprint*, arXiv:2204.13601, 2022.
[14] A. Yazdani and Y. Shekofteh, ”A Persian ASR-based SER: Modification of Sharif Emotional Speech Database and Investigation of Persian Text Corpora,” *arXiv preprint*, arXiv:2211.09956, 2022.
[15] A. Yazdani et al., ”Persian Speech Emotion Recognition by Fine-Tuning Transformers,” *arXiv preprint*, arXiv:2402.07326, 2024.
[16] F. Sarlakifar, ”Persian Text Emotion Recognition by Fine-Tuning the XLM-RoBERTa Model,” *GitHub Repository*, 2023.
[17] Rasouli et al., ”ShortPersianEmo Dataset,” *Papers with Code*, 2021.
[18] ”Macro-Average: Rare Types Are Important Too,” *Data Science Articles*, 2023. Available at: https://datasciencearticles.com/macro-average-importance.
[19] S. Gururangan et al., ”Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks,” *arXiv preprint*, arXiv:2004.10964, 2020.
[20] H. Chang and H. Lu, ”Exploring Intermediate Fine-Tuning for Domain Adaptation of Pretrained Models,” *ACL Anthology*, 2021.
[21] I. Alghanmi, M. Elmadany, and D. Khashabi, ”Self-Supervised Intermediate Fine-Tuning of Biomed- ical Language Models for Interpreting Patient Case Descriptions,” *ACL Anthology*, 2022.
[22] F. Sarlakifar, M. Mahdavi Mortazavi, and M. Shamsfard, ”EmoRecBiGRU: Emotion Recognition in Persian Tweets with a Transformer-based Model, Enhanced by Bidirectional GRU,” *International Journal of Information and Communication Technology Research*, vol. 16, no. 3, pp. 35–44, 2024. Available: https://ijict.itrc.ac.ir/article-1-653-en.pdf.