ArmanEmo: A Persian Dataset for Text-based Emotion Detection

Hossein Mirzaee
Department of Chemical Engineering
Amirkabir University of Technology
Tehran, Iran
hmirzaee@aut.ac.ir

Javad Peymanfard
School of Computer Engineering
Iran University of Science and Technology
Tehran, Iran
javad_peymanfard@comp.iust.ac.ir

Hamid Habibzadeh Moshtaghin
Faculty of Management and Accounting
Allameh Tabataba'i University
Tehran, Iran
h.habibzadeh@atu.ac.ir

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

Abstract

With the recent proliferation of open textual data on social media platforms, Emotion Detection (ED) from Text has received more attention over the past years. It has many applications, especially for businesses and online service providers, where emotion detection techniques can help them make informed commercial decisions by analyzing customers/users' feelings towards their products and services. In this study, we introduce ArmanEmo, a human-labeled emotion dataset of more than 7000 Persian sentences labeled for seven categories. The dataset has been collected from different resources, including Twitter, Instagram, and Digikala [1] comments. Labels are based on Ekman's six basic emotions (Anger, Fear, Happiness, Hatred, Sadness, Wonder) and another category (Other) to consider any other emotion not included in Ekman's model. Along with the dataset, we have provided several baseline models for emotion classification focusing on the state-of-the-art transformer-based language models. Our best model achieves a macro-averaged F1 score of 75.39 percent across our test dataset. Moreover, we also conduct transfer learning experiments to compare our proposed dataset's generalization against other Persian emotion datasets. Results of these experiments suggest that our dataset has superior generalizability among the existing Persian emotion datasets. ArmanEmo is publicly available for non-commercial use at https://github.com/Arman-Rayan-Sharif/arman-text-emotion.

1 Introduction

Emotion expression and detection play important roles in human social and professional life. They are closely related to our cognitive abilities and communication skills, profoundly shaping the range and quality of our experiences and social interactions [1]. Consequently, emotional intelligence is one of the essential abilities to move from narrow to general human-like Artificial Intelligence [2], [3]. Emotion Detection (ED) is an active area of research to enable machines to understand human emotions effectively.

Emotion detection and analysis have been widely researched and have many applications in several fields such as neuroscience, psychology and behavioral sciences, computer science, and computational linguistics [4], [5]. In business and online commercial activities, companies and organizations are eager to analyze their customer's emotions toward the company and their products so they can adapt their behavior, services, and products to the customer expectations and hence, ensure the business growth [6]. Due to the current expansion of the social platforms and the recent proliferation

[1] an Iranian e-commerce company.

of open conversational data, Emotion Recognition in Conversation (ERC) has become another attractive subfield of ED, which in turn has many potential applications in important downstream tasks such as emotionally intelligent chatbots, emotion-aware health-care, and educational systems, opinion mining and recommender systems [3, 7, 8, 9, 10, 11].

Emotions can be expressed verbally (through words and tone of voice) or nonverbally (facial expressions or body language), and they, therefore, can be recognized in many sources like images and videos [12], sounds [13], texts [14]. With respect to ED from voice/speech, images, videos, and other multimodal data, ED from text is more challenging and has been studied less than the other approaches [6]. This is because the text alone may be insufficient to recognize emotions [15]. Even for humans, knowing the context of a conversation is essential to recognize the emotions from a text, given the lack of other data such as tone of voice and facial expressions in this case [8]. Diversity in the style of writing, formal and informal styles [16], the usage of figurative language, like sarcasm [17], detecting emotions from short texts, and presence of grammatical errors [6] are some of the existing difficulties, which make the ED from text more challenging than it might seem at first glance. Regardless, the ever-increasing rise in the amount of available textual data in social media and its vast potential applications have made the ED from text an important domain to study [8].

The goals of this paper are to:

- Conduct a review study on Text-based Emotion Detection (especially Emotion Detection from Persian Text)
- Contribute to Persian Text ED:
  - By introducing a new dataset for Emotion Detection
  - By introducing strong baseline language models

The rest of this paper is organized as follows: in section 2 the models of emotion and different approaches to emotion detection from text have been discussed, and a few studies regarding emotion detection from Persian text has been reviewed. section 3 provides our main contributions to emotion detection from Persian Text ED by introducing ArmanEmo, our new dataset, and the procedures we have followed to generate it. This section also presents several baseline models trained on ArmanEmo and their performance, along with various experiments to demonstrate the high quality of the dataset. Finally, section 4 concludes the paper.

2 Text Based Emotion Detection

In this section, we first discuss emotion models used in previous studies related to Emotion Detection. We then review different approaches used to detect emotion from text. Finally, in the last part, we provide a brief review of previous works in emotion detection from Persian text.

2.1 Emotion Models

Fundamental to the development of Emotion Detection systems are emotion models that determine how various human affects are represented. Affect studies have been done in a variety of fields such as neuroscience, psychology, and cognitive sciences. According to research in psychology, there are three distinguished approaches for emotion modeling which are Categorical, Dimensional, and Appraisal-based approaches [18], [19]. However, the first two are the most important and often used approaches in emotion detection studies [6, 4].

The categorical approaches suggest that human beings have a limited number of psychological and biologically basic emotions that are universally recognized [20]. So, they involve classifying emotions into distinct classes or categories. The most commonly adopted approach within the categorical or discrete approaches is that of Paul Ekman [21], which proposes that the six basic human emotions are ANGER, DISGUST, FEAR, HAPPINESS, SADNESS, and WONDER. Plutchik adds two more emotions to Ekman's set of emotions, meaning TRUST and ANTICIPATION, so the whole set can be organized into four bipolar subsets: joy vs. sadness, anger vs. fear, trust vs. disgust, and surprise vs. anticipation [22].

While the majority of the automatic Emotion Systems have been based on categorical approaches, some researchers argue that any small set of basic emotion classes may not reflect some of the non-basic and complex affective states of human communications like thinking or depression [23]. Hence, they suggest representing emotions in a dimensional form. The main idea is that since the emotions are not independent, we need to systematically capture their relations with each other by placing them in a 1D, 2D, or 3D spatial space [19]. The most widely used dimensional model is Russell's Circumplex Model of Affect [24], suggesting that emotions are distributed in a two-dimensional circular space:

- Valence dimension: which differentiates emotions by Pleasantness and Unpleasantness
- Arousal dimension: which indicates how excited or apathetic an emotion is

Most computational approaches for Emotion Detection are based on categorical emotion models, mainly because of their simplicity and familiarity. Nonetheless, since they are limited to a fixed number of emotions, they may not cover non-basic, mixed and complex emotional states [25]. On the other hand, Dimensional approaches are more helpful in depicting these subtle emotions that differ only slightly. They are also highly recommended when the goal is to measure similarities between emotions [6]. However, fitting all basic emotions in the dimensional space is not feasible since some become indistinguishable, and some may lie outside the space [19]. As we can see, both approaches have advantages and disadvantages, and none of them is superior to another in all situations. The selection of an emotion model depends on the end goal of developing an Emotion Detection system and the set of emotions we expect from the system to detect [4].

2.2 Emotion Detection Approaches

Among different computational approaches for Emotion Detection from text, three of them are currently more dominant: rule-based, Machine Learning-based, and hybrid approaches [26]. The rule-based approaches are based on following grammatical and logical rules to classify the text into different emotion categories. Such rules can be determined using linguistic, statistical, or computational concepts. One of the most widely used rule-based approaches is keyword recognition which relies on finding occurrences of predefined keywords in a given text at the sentence level. This keyword list or dictionary is prepared with the semantic labels of emotion such as sadness or anger, and sometimes it also indicates the intensity of the emotion. Once the keyword is identified within the sentence, emotion intensity measurement and negation checking are performed, and finally, an emotional label is assigned to the sentence [16]. The keyword-based approaches are straightforward and intuitive, yet, they face some challenges. Limited and domain-specific keyword lists and poor pre-processing heavily affect the emotion detection performance [6, 27].

In ML-based approaches, trained classifiers are used to automatically assign an emotional label to the input sentence. Supervised learning approach is one type of ML-based technique that relies on extensive training data annotated with emotional tags or labels. Trained on the training set to learn how to classify a given text into an emotion label, the supervised classification algorithm infers a function, which eventually can be used to map any new unseen examples into emotion labels. Unlike keyword-based approaches, supervised learning approaches are more adaptable to changes in the domain since they can effectively and quickly learn new features from the text [26]. However, to train such classifiers, a large labeled dataset is needed, which may be a tedious and time-consuming task.

Among classical supervised learning algorithms, SVM has been widely used for Emotion Detection from text [28]. Given the meaningful feature set (obtained after performing some preprocessing, linguistic and statistical analysis on the input texts, and emotion labels), the SVM algorithm outputs an optimal hyperplane in multidimensional space to separate different emotion labels. Deep Learning algorithms are the other supervised ML-based approaches that have recently received considerable attention in Emotion Detection from text. It is argued that the deep layers of these algorithms enable them to capture the variations in the meaning of a word depending on its context [29]. Different variations of Recurrent Neural Networks, such as LSTM and GRU, Convolutional Neural Networks, and Transformer-based models are some techniques that have been used in Emotion Detection from text [16]. Implementation of some of these architectures has led to state-of-the-art results in Emotion Detection [6].

Unsupervised learning algorithms are the other type of ML-based approaches. In unsupervised learning approaches, instead of training a model on labeled or annotated data, the goal is to find some hidden structures in unlabeled data upon which models for emotion detection can be built [4].

In hybrid approaches, rule-based approaches and ML-based approaches are consolidated into a new solution that has the strengths of both approaches while alleviating their associated weaknesses. The idea is that one can yield more accurate results in emotion detection by implementing an ensemble of classifiers and adding knowledge-rich linguistic information from dictionaries [26]. It should be noted, however, that the performance enhancement a hybrid method can provide heavily relies on the particular types of classification methods used [6].

2.3 Previous works in Emotion Detection from Persian Text

Very few studies have been conducted on emotion detection from Persian text. In one of these studies [16], a hybrid approach, based on the combination of cognitive features and Word2Vec embeddings, is proposed to achieve better performance in emotion detection. The emotional constructions, keywords, and parts of speech are utilized to construct the cognitive features. The resulting embedding vector is then used in a GRU network. It is suggested that this hybrid method outperforms the methods that rely solely on linguistic information or deep learning approaches. However, in general, the linguistic rules cannot be used in many practical applications. For instance, it is challenging to extract cognitive features in social media where the usage of colloquial and informal language is dominant. It is noteworthy that, to perform the experiments, the authors have sampled 23,000 sentences from Bijan Khan's corpus, making sure that each sentence carries only one emotion among five basic emotions (fear, happiness, sadness, anger, and surprise). Unfortunately, this dataset is not publicly available.

In another study, EmoPars, a dataset containing 30,000 Tweets, has been published [30]. EmoPars is the largest source of Persian text-based emotions; however, it has problems that make this large amount of labeled data inefficient. It is labeled through crowd-sourcing, in which each sentence has been given to 5 voters, and the final labels are to be acquired using maximum voting. Moreover, a significant number of selected tweets do not contain any emotions and are neutral. Only 23% of the total dataset, including 7026 samples, are labeled with emotions. To evaluate the quality and generalizability of this dataset against ArmanEmo, we perform an experiment, comparing the result of the best model in our baselines when it is trained on Emopars against when it is trained on our dataset. This experiment is described in detail in section 3.3.3.

3 Our Study

This section presents our contributions toward emotion detection from Persian text. First, we discuss the procedures we have followed to collect and annotate our Emotion dataset. Statistics regarding ArmanEmo are also provided. In the second part of this section, we introduce our baseline models and describe the preprocessing steps we have followed. Finally, in the third part, we present and discuss the results of all the experiments.

3.1 Our New Emotion Dataset

In this section, we demonstrate our new emotion dataset along with its statistics and describe the methodology we have followed for data collection and annotation.

3.1.1 Data Collection

We have used different resources to collect the raw textual data we needed for this study. Of the most important resources that are recently receiving more attention in the field of emotion classification studies is the large text corpus coming from social media platforms. This is because individuals are increasingly using these online platforms to communicate their ideas and feelings about various topics. Since one of the main goals of this project is to evaluate the emotions towards different social and political topics, we have included Persian tweets published on Twitter, one of the widely used social media platforms in Iran. However, to make ArmanEmo more general and representative, we have also used two other resources. Along with a text corpus from users' comments on Instagram, we have included customers' comments on Digikala (an online shopping platform) in the dataset. Table 1 summarises the resources we have used and the related details about them.

Table 1: Resources used to collect textual data

	Persian Tweets	Instagram Comments	Digikala Comments
Collection Method	Tweeter's official API	Facebook Graph API	Polite Crawling of the Website
Collection Period	2017 - mid 2018	mid 2017 - mid 2018	mid 2018
# Raw Data	1.5 M	1 M	50 K
# Data Used for Manual Annotation	3.5 K	3 K	1 K
# Data Used for Automatic Annotation	4.5 K	-	-

3.1.2 Data Annotation

After data collection, we took the following steps to improve the data quality before the annotation process:

1. For the data coming from each of the above-mentioned resources, after calculating the distribution of sentences according to their lengths (in characters), we removed all the sentences whose lengths felt outside a specific range. We also eliminated sentences containing specific user IDs or links.

2. In order to loosely control the class balance in our target dataset, we used a heuristic technique to weakly label each given sentence in the dataset before presenting them to our annotators for labeling. These "weak labels" are not final and are not guaranteed to be consistent with the labels selected by our annotators (weak labels were hidden from our annotators to avoid any bias in the labeling system). However, they are still helpful tools during the labeling process to pre-select sentences from rare classes more often so that the resulting dataset gets as much balanced as possible. The heuristic technique we used to generate weak labels relies on the NRC Word-Emotion Association Lexicon [31], a dataset containing a list of emotional words and their association with basic emotions. We pre-classified the sentences coming from each of the three resources according to the association between their words and each of the emotion classes. We also used a weighted random selection method based on the Term-Frequency of emotional words in the sentences to make sure that those sentences containing emotional words with higher Term-Frequency are more likely to be selected and shown to the annotators during the labeling process.
3. In the end, we chose 12000 sentences to be labeled in the next step. Before handing these to our annotators (in the manual annotation process), we removed any emojis in these sentences to make sure they categorize the emotions in each given sentence only based on its text.

Here we describe the annotation procedure, which is a mix of manual and automatic steps. Manual data annotation was performed on 7500 sentences (out of 12000 selected sentences) through the application of a Telegram Bot developed for this project using tools like official Telegram APIs and MySQL. At the beginning of their interaction with this bot, users were first introduced to our specially designed instructions for data annotation. Then, their performances in following our rules and instructions were evaluated and measured against our existing labeled test set. The test set containing 250 sentences had already been manually labeled and validated by ten annotators. Among the 35 users who participated in the data labeling, we selected 12 of them with the highest scores as our final annotators. The data annotation was done in March 2019.

Based on our instructions, given a sentence, annotators have a few options to select. If annotators are sure that the sentence has only one specific emotion among our six predefined categories, they would need to choose that emotion as the label of the sentence. If they decide that the sentence carries no specific emotion or emotions other than our predefined classes, they will select "Other." Otherwise, if they cannot confidently assign a specific feeling to the given sentence, they should select the "Unknown" option.

By the end of the labeling loop, each given sentence needed to be classified into the same emotion category by three different annotators considering that it had not been shown to more than five annotators. In other words, if the sentence had already been shown to five annotators without being finalized, it would be removed from the labeling procedure and will not be shown to other annotators until the last step of the labeling (to be explained later). To speed up the labeling process, we did not hand the sentences to the annotators all at once. Instead, we prioritize the introduction of the sentences to the annotators based on their previous labels so that those sentences that needed less labeling effort to be finalized (and temporarily removed from the labeling loop) get displayed to the annotators sooner than the others. For example, a sentence labeled by three annotators without being finalized is of more priority with respect to the sentence that has been classified once (by only one annotator).

In order to control the quality of labels during the labeling process, we also used a data dashboard to capture and display the instant information related to our annotation system. The aim for the development and application of such a system was:

1. **To balance the class distribution in the final labeled dataset**: Observing the class distribution of the labeled data during the annotation process, we tried to update the selection weights (by which new weakly-labeled sentences were introduced to our annotators) in reverse proportion to the frequency of each label in our labeled data. We wanted to present new sentences with rare labels (i.e., their corresponding weak-label) to our annotators more often so that we have a final dataset that is as much balanced as possible. It is worth noting that the weak labels are still weak, and they were likely to be rejected by our annotators. Due to this reason, our final labeled dataset is not perfectly balanced.
2. **To analyze the performance of the annotators**: During the annotation process, we evaluated the performance of each annotator based on their contribution to labeling the finalized examples. The more they had involved in finalizing the labels for input data, their performance score was higher. Among the twelve annotators, three of them got the highest score whose judgment and skills were used in labeling the remaining examples that had not been finalized in previous steps. One of the annotators, on the other hand, had a very low and unacceptable score and was eliminated from the annotation step.

By the end of the manual annotation steps, we had 4700 sentences labeled by our annotators. We randomly split this data into primary training (>3500 samples) and testing sets (>1100 samples). We then fine-tuned ParsBert Language

Model [32] on this training set (with the same hyperparameters described in section 3.2). We used this classifier to label the second parts of our selected input sentences (4500 samples from Persian Tweet, 1). Among these sentences, 2000 samples with lower confidence were filtered out from our annotation process. An annotator then manually checked the labels of the remaining 2500 sentences and changed them whenever needed. These 2500 samples were then added to our training set.

3.1.3 Dataset Statistics

As mentioned before, we selected 7500 sentences to be manually labeled as one of the emotion classes or two other categories, i.e., "Unknown" and "Other." Figure 1 shows the count of labeled and unlabeled data per attempt (each iteration of the annotation process). We observed that the labels of 38 percent of these 7500 sentences ended up undecided because they did not label as one of the emotion classes after reviewing by the fifth annotator. This observation is another piece of evidence showing that emotion classification from text is a challenging task even for human annotators.

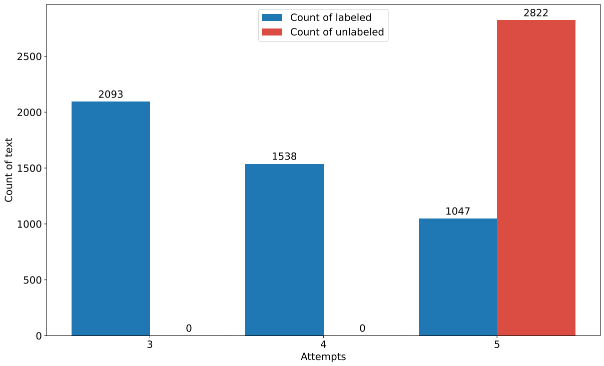

Figure 1: Count of labeled and unlabeled data per attempts

In figure 2 the label frequency for all the (manually and automatically) labeled data is provided. Twenty-five percent of the final labeled sentences had been classified as "Other." Moreover, some data was labeled as "Unknown" by the end of the annotation process, which we decided to remove from the dataset.

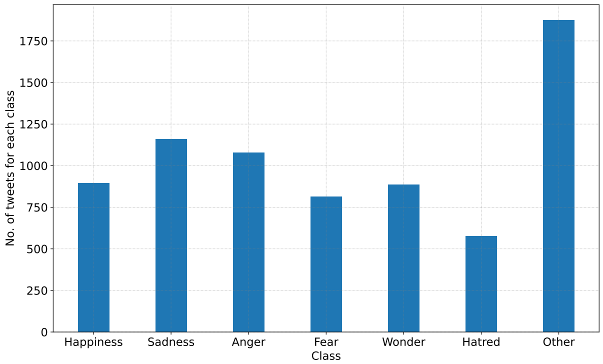

Figure 2: No. of Tweets for Each Class

3.2 Modeling

In this section, we provide baseline models for emotion classification over ArmanEmo, which use transfer learning. The fact is that supervised learning is difficult to apply to NLP problems, including emotion detection, since labels are costly. Here is where transfer learning comes into play. Transfer learning from pre-trained deep neural Language Models (LMs) towards downstream language problems has led to a state-of-the-art performance in several NLP tasks in recent years [33, 34, 35]. Deep LMs can be efficiently pre-trained in an unsupervised way over very large unlabeled datasets. In this way, the resulting LMs will capture rich and non-trivial linguistic knowledge, making them suitable to be transferred to a subsequent target domain and task through supervision. To be adaptive to a target domain, the pre-trained LMs need to be fine-tuned by a small amount of labeled data from that domain.

As one of our baseline models, we take advantage of a pre-trained language model for Persian, known as ParsBERT. ParsBERT is a monolingual language model based on Bidirectional Encoder Representation Transformer (BERT) architecture. Farahani et al. [32] have shown that the ParsBERT model outperforms the multilingual BERT and previous models in several Persian NLP downstream tasks, including text classification and sentiment analysis. Lighter than the multilingual BERT, ParsBERT has been trained on a larger and more diverse (in terms of the range of topics and style of writing) set of pre-trained Persian datasets.

We also use two variations of a model known as XLM-RoBERTa as our other baseline models. XLM-RoBERTa is another transformer-based masked language model pre-trained on text in 100 languages. This multilingual language model has led to state-of-the-art performance on cross-lingual classification, sequence labeling, and question answering, outperforming multilingual BERT (mBERT) on various cross-lingual benchmarks [36]. Although it is known that ParsBERT as a monolingual language model outperforms multilingual BERT on various tasks in Persian language, we decided to compare the performance of XLM-RoBERTa variations (namely XLM-RoBERTa-base and XLM-RoBERTa-large) against ParsBERT on emotion detection from Persian text.

Another model included in our baselines is XLM-EMO [37], a multilingual emotion detection model for social media text. It is essentially XLM-T fine-tuned on datasets for emotion detection in 19 different languages. XLM-T itself is a fine-tuned version of XLM-RoBERTa-base on Twitter data [38]. In developing XLM-EMO model, emotion labels of each dataset are mapped to a common set, namely joy, anger, fear, and sadness. It has been shown that this multilingual emotion model is competitive against language-specific baselines in zero-shot settings. It is specifically developed to help low-resource languages that still do not have a dataset for emotion detection. In our study, we evaluate the performance of this model in three different settings. In the zero-shot setting, we want to assess this model's accuracy in emotion detection from Persian text without being trained on our emotion dataset. To have a basis of comparison for the zero-shot learning in the first setting, we also fine-tune XLM-EMO on our train set and evaluate it on our test set. To perform these two studies, we filter out the missing emotions from ArmanEmo so that the models can predict only joy, anger, fear, and sadness. Moreover, to compare the performance of XLM-EMO to other models in our baselines, we also train and evaluate another XLM-EMO model considering all seven emotion classes in our original dataset.

To fine-tune the pre-trained language or emotion models for our task, i.e., Emotion Detection, we add a fully-connected dense layer on top of these models. We also introduce a cross-entropy loss function to perform the multi-label classification task using the resulting models. While fine-tuning, we freeze all the networks' weights except the weights for the last layer, i.e., the added dense layer.

3.2.1 Data Pre-processing

After splitting the data into training and testing sets, we follow some pre-processing steps to transform the data into a format proper to feed into the final model. As the first pre-processing step, we normalize the text using Parsivar [39], a toolkit for Persian text pre-processing. It applies some rule-based space correction steps (including word, punctuation, and affix spacing), along with some character refinement operations (such as removing stretching letters). Its normalization rules, however, are not comprehensive. For example, "Arabic Sukun" will not be removed after normalization. Besides, we need to consider some task and domain-specific rules while pre-processing the text from social media. So, after introducing the text to the Parsivar normalizer, we perform some additional pre-processing steps, which include removing:

1. any English character from the text

2. letters repeated more than twice in the non-standard spelling of Persian words which are often intentionally used for more emphasis in the informal text (like خیییلللي instead of خیلی)

3. any Arabic diacritics from the text which are not removed by Parsivar normalizer

4. any remaining non-Persian characters after performing the above steps
5. the hashtag sign ("#") from the text while keeping the information included in the hashtags
6. Persian numeric characters from the text

3.2.2 Hyperparameters

When fine-tuning the pre-trained ParsBERT model, we used the same hyperparameters set by Farahani et al. except for the batch size, maximum sequence length, and the number of epochs. We fine-tuned the ParsBERT on our training dataset for eight epochs using a batch size of 32 while limiting the maximum sequence length to 128. We used AdamW (Adam with decoupled weight decay) optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\lambda = 0.01$. We also utilize a learning rate scheduler, linearly decreasing from the initial learning rate (2e-5) to 0 by the end of the last epoch. For other Language Models, we have utilized the default hyperparameters used in the open-source library Transformers [40].

3.3 Results

In this section, we discuss the results of different experiments performed to showcase the quality of ArmanEmo. The first part of this section summarizes the performance of various deep models on ArmanEmo using different evaluation metrics. The results for the Zero-shot tests using XLM-Emo on the dataset are discussed next. Lastly, we compare the generalization of ArmanEmo against EmoPars in a transfer learning setting.

3.3.1 Comparison of baseline models on ArmanEmo

In order to compare the emotion models based on fine-tuned Language Models in our baselines against other deep neural networks, we trained and evaluated CNN-based and RNN-based models on our train and test datasets. It is worth noting that for all the models that are not based on language models, we used the same pre-trained word vectors as the embeddings, which are trained on Common Crawl and Wikipedia using fastText [41]. As Table 2 shows, the fine-tuned XML-RoBERTa-large model significantly outperforms other models on our test set in terms of average macro F1 score, precision, and recall. Moreover, it can be seen that there is a performance gap between fine-tuned LMs and RNN/CNN-based models in emotion detection. This significant performance difference can be related to the superior ability of pre-trained Language Models in extracting rich and non-trivial knowledge from the textual data.

deep neural netwroks

Table 2: Comparison between the performance of different DNN models and Language Models

Model	Precision (Macro)	Recall (Macro)	F1 (Macro)
FastText [42]	54.82	46.37	47.24
HAN [43]	49.56	44.12	45.10
RCNN [44]	50.53	48.11	47.95
RCNNVariant	51.96	48.96	49.17
TextAttBiRNN [45, 46]	54.66	46.26	47.09
TextBiRNN	51.45	47.16	47.14
TextCNN [47]	58.66	51.09	51.47
TextRNN [48]	49.39	47.20	46.79
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

The performance of the best model among our baseline models (XLM-RoBERTa-large) is summarized in Table 3 and Figure 3. The model achieves a macro average F1 score of 75.39 on the test set. It presents the best performance on emotions like Happiness, Fear, and Sadness. On the other hand, it obtains the lowest F1 score on emotions like Anger, Other, and Hatred.

Table 3: Evaluation Metrics Resulted from The Best Model (XLM-RoBERTa-large)

Emotion	Precision	Recall	F1	Support (No. of Test Examples)
Anger	74.62	62.99	68.31	154
Fear	78.69	84.21	81.36	57
Happiness	86.02	87.27	86.64	275
Hatred	63.64	75.38	69.01	65
Other	60.98	77.72	68.34	193
Sadness	82.45	77.10	79.68	262
Wonder	84.96	66.21	74.42	145
Macro Average	75.91	75.84	75.39	1151

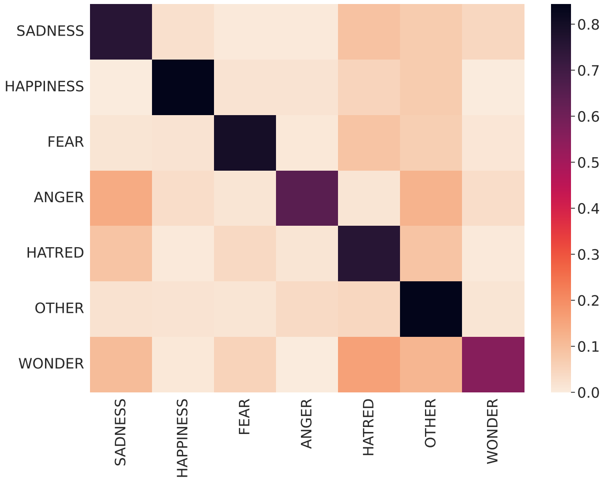

Figure 3: Confusion Matrix for XLM-RoBERTa-large Predictions

We manually investigated the mislabeled sentences to better analyze the situations where the best model is performing poorly. Table 4 presents some randomly selected samples of these mislabeled sentences generated by the best model. Going through some of these examples, what can be inferred is that the model outputs the wrong label whenever a given sentence carries mixed emotions. In such situations, the assignment of one and only one exact emotion to the sentence might be challenging even for humans. That is why multi-label classifiers are used against multi-class classifiers to include the presence of more than one emotion in a given sentence. Since we are dealing with a multi-class classification problem in this study, we presume that each sentence carries only one emotion; hence, it is assumed that there is only one true target for each sentence. However, this might not be the case for all situations. One can judge, based on the given sentences, that some of the model's predictions are not irrelevant at all. In fact, depending on the context, these predicted labels might be considered as valid as the assigned ground truths. However, there are still other situations where the model's poor performance seems to be related to the model being biased to the occurrence of some specific words or combinations of words in the sentence.

Table 4: Randomly-selected samples of mislabeled sentences (generated by XLm-RoBERTa-large)

Sentence	Ground Truth Label	Model Prediction
واقعا حال به هم زنه این حجم از داستان سرایی درباره تجاوز یا چیزهاي شبیه به اون برا جذب لایك و توجه The amount of fake stories being spread about rape or things like that for likes and attention is honestly revolting.	Hatred	Wonder
کتاب امروز به دستم رسید. جنس برگه هاش خوب بودن. رنگهای شادی داشت. اسم شخصیتهای داستانها خنده دار و جالب بود. داستانهاش هم تکراری نبودن. فقط چون کل سی جلد یکجا جمع شده کمی دست رو خسته میکنه I got my hands on the book today. The paper has good quality. A lot of vibrant colors. The characters in each story had funny and interesting names. The stories weren't repetitive. Although because it was a collection of all 30 books your hands would get tired.	Happiness	Other
انصافا چقدر فک کردی اینو نوشتی خخخ! ولی احسنت جالب بود Honestly how much thought did you put in when you wrote this, LOL! But nice, it was interesting	Happiness	Wonder
بریم یک سیگاری بکشیم شاید حال داد. Lets go smoke a cigarette, it might be fun.	Happiness	Other
چجوری میشه آدم خوشحالو خندون باشه بعد یه دفعه غمگین بشه؟ فکنم دچار نوع خاصی از خود درگیری شدم How can a person be happy and suddenly feel so down? I think I am struggling with a specific kind of internal conflict.	Sadness	Wonder
بزرگترین گناه ترس هست و این خانم شجاعترین مرد میدان بود درود بر شرفش The biggest sin is fear and this woman was the bravest hero. Peace be upon her.	Happiness	Fear
سلبریتیا میان از وضعیت بد اقتصادی مردم انتقاد میکنن بعد بلیط تئاتر خودشون صد و هشتاد هزار تومنه :)) The celebrities want to complain about the bad economic times yet they make their own theater shows cost 180 Tomans :))	Wonder	Anger
موسیقی و شاد بودن حرام نیست Listening to music and being happy isn't haram.	Other	Happiness
در مورد راه شیری چیز خاصی نمیشه گفت حیرت انگیزه! ولی از خفن بودن تلسکوپ هابل هم نمیشه گذشت I cant find the words to describe the Milky Way since it is truly breathtaking! But that also doesn't take away from how amazing the Hubble telescope is.	Wonder	Happiness
ملت رفتن زیر آخرین پستهای اینستاگرام مرحوم گفتن روحت شاد. Everyone's gone under the decedent's most recent post on Instagram and have said rest in peace.	Wonder	Sadness

3.3.2 Zero-shot tests using XLM-EMO

Table 5 compares the performance of XLM-EMO (a multilingual emotion model) as a zero-shot classifier against another version of it which is trained on our dataset (while the data is limited to only four basic emotions, as is explained in section 3.2) (the F1 score for the XLM-RoBERTa-large model and ParsBERT are also provided for comparison). As expected, the model trained on our train set has a superior performance in terms of macro-averaged F1 score. However, the performance of XLM-EMO as a zero-shot classifier, which has not seen any of our Persian text during its training, is still impressive. It's performance is somewhat comparable to the performance of ParsBERT, which has been pre-trained on large Persian corpus. This result suggests that using a multilingual emotion model can be very helpful in detecting emotion from Persian text in the lack of any emotion dataset.

Table 5: Result for zero-shot test using XLM-EMO while considering four basic emotions

Model	F1 Score
XLM-EMO Zero-Shot	75.28
XLM-EMO Trained	84.46
XLM-RoBERTa-large	85.02
ParsBERT	79.38

3.3.3 Comparing emotion datasets for Persian text

To better demonstrate the capabilities of our emotion dataset, we compare the performance of the best model in our baselines (XLM-RoBERTa-large) when it is trained on ArmanEmo against when it is trained on EmoPars. Before this experiment, some preparatory steps need to be done on EmoPars since the sentences in this collection are released with no major emotion specified by the authors. Instead, Sabri et al. have directly published the results of their crowd-sourced data labeling procedure. Each sample in EmoPars contains a sentence along with six numbers (vote counts in the range of [0,5]) corresponding to each of the six emotions considered in this study. Selecting the final emotion for each sentence can be done by introducing a threshold, a task delegated to the users of EmoPars. For our purposes, we decide to remove samples with no dominant emotion, and, to do this, we define a dominance threshold which is set to 3. In other words, we filter out the sentences for which none of the six emotions has received a YES vote from 3 out of 5 annotators. For the remaining samples, if the sentence has more than one dominant emotion, it gets removed again. Otherwise, the dominant emotion will be selected as the final label of the given sentence. Following this procedure, we selected 5477 tweets (out of 30000 samples) from EmoPars. Since the original dataset is not split into train and test sets, we randomly selected 80 percent of these tweets as the train set and the remaining 20 percent as the test set.

To make a fair comparison, we run our experiments in four different situations. In the first two combinations, we use XLM-RoBERTa-large, which is trained on our train set, and in the last ones, we use XLM-RoBERTa-large, which is trained on the train set of EmoPars. For both cases, we evaluate the model's performance on our test set and on the test set of EmoPars. The results of this study (in terms of macro-averaged F1 score) are summarized in Table 6

Table 6: Macro-averaged F1 score results for comparison between ArmanEmo and EmoPars

Train Set	Tested on EmoPars	Tested on ArmanEmo
ArmanEmo	**26.68**	**75.39**
EmoPars	6.96	7.16

As it can be seen in Table 6, the model that has been trained on our train set has a superior performance to the model trained on the train set of EmoPars, even when the models are evaluated on the test set of EmoPars. When testing both models on the test set of EmoPars, the F1 score is more than 19 percent higher for the model trained on ArmanEmo. This observation indicates that the generalizability of ArmanEmo is superior to that of EmoPars.

The noticeably poor performance of the models that were trained on EmoPars can be related to the fact that this dataset is labeled through crowd-sourcing, consisting of many noisy labels. In our study, on the other hand, the labels were selected according to a meticulously designed procedure discussed in section 3.1. We have chosen the labels with much fewer errors, so it was expected that ArmanEmo would have less noise and, therefore, better generalizability.

4 Conclusion

In recent years, emotion detection from text has been a topic of interest to researchers in natural language processing. Among different approaches to developing emotion detection systems, data-driven methods in general, and deep learning algorithms in particular, have received considerable attention in this field. A significant issue with deep learning algorithms in emotion detection from text is the lack of annotated labels. In this study, we provide a manually annotated dataset suitable for training deep learning algorithms. To demonstrate the high quality of our dataset, we build several strong baseline models, including state-of-the-art language models. Through transfer learning experiments, we show the superior generalizability of our proposed dataset. The final dataset, containing more than 7000 samples, is publicly available for non-commercial use.

Acknowledgement

The authors would like to thank Arman Rayan Sharif for providing required financial and computational resources in this project.

References

[1] Raymond J Dolan. Emotion, cognition, and behavior. *science*, 298(5596):1191–1194, 2002.

[2] Waleed Ragheb, Jérôme Azé, Sandra Bringay, and Maximilien Servajean. Attention-based modeling for emotion detection and classification in textual conversations. *arXiv preprint arXiv:1906.07020*, 2019.

[3] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953, 2019.

[4] Lea Canales and Patricio Martínez-Barco. Emotion detection from text: A survey. In *Proceedings of the workshop on natural language processing in the 5th information systems research working days (JISIC)*, pages 37–43, 2014.

[5] Carlo Strapparava and Rada Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560, 2008.

[6] Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189, 2020.

[7] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[8] Chenyang Huang, Amine Trabelsi, and Osmar R Zaïane. Ana at semeval-2019 task 3: Contextual emotion detection in conversations through hierarchical lstms and bert. *arXiv preprint arXiv:1904.00132*, 2019.

[9] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825, 2019.

[10] Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687*, 2019.

[11] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*, 2018.

[12] Hadjer Boubenna and Dohoon Lee. Image-based emotion recognition using evolutionary algorithms. *Biologically inspired cognitive architectures*, 24:70–76, 2018.

[13] Akshay Chatterjee and Ghazaala Yasmin. Human emotion recognition from speech in audio physical features. In *Applications of computing, automation and wireless systems in electrical engineering*, pages 817–824. Springer, 2019.

[14] Fabio Calefato, Filippo Lanubile, and Nicole Novielli. Emotxt: a toolkit for emotion recognition from text. In *2017 seventh international conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 79–80. IEEE, 2017.

[15] Chenyang Huang, Osmar R Zaiane, Amine Trabelsi, and Nouha Dziri. Automatic dialogue generation with expressed emotions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 49–54, 2018.

[16] Seyedeh S Sadeghi, Hasan Khotanlou, and M Rasekh Mahand. Automatic persian text emotion detection using cognitive linguistic and deep learning. *Journal of AI and Data Mining*, 2021.

[17] Ankush Chatterjee, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, 93:309–317, 2019.

[18] Didier Grandjean, David Sander, and Klaus R Scherer. Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. *Consciousness and cognition*, 17(2):484–495, 2008.

[19] Hatice Gunes and Maja Pantic. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)*, 1(1):68–99, 2010.

[20] Paul Ekman and Wallace V Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*, volume 10. Ishk, 2003.

[21] Tim Dalgleish and Mick Power. *Handbook of cognition and emotion*. John Wiley & Sons, 2000.

[22] Robert Plutchik. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984:197–219, 1984.

[23] Hatice Gunes, Björn Schuller, Maja Pantic, and Roddy Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 827–834. IEEE, 2011.

[24] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[25] Chen Yu, Paul M Aoki, and Allison Woodruff. Detecting user engagement in everyday conversations. *arXiv preprint cs/0410027*, 2004.

[26] Haji Binali, Chen Wu, and Vidyasagar Potdar. Computational approaches for emotion detection in text. In *4th IEEE International Conference on Digital Ecosystems and Technologies*, pages 172–177. IEEE, 2010.

[27] Jeffrey T Hancock, Christopher Landrigan, and Courtney Silver. Expressing emotion in text-based communication. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 929–932, 2007.

[28] Nourah Alswaidan and Mohamed El Bachir Menai. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge & Information Systems*, 62(8), 2020.

[29] Yen-Hao Huang, Ssu-Rui Lee, Mau-Yun Ma, Yi-Hsin Chen, Ya-Wen Yu, and Yi-Shin Chen. Emotionx-idea: Emotion bert–an affectional model for conversation. *arXiv preprint arXiv:1908.06264*, 2019.

[30] Nazanin Sabri, 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*, pages 167–173, 2021.

[31] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[32] Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*, 53(6):3831–3847, 2021.

[33] Neel Kant, Raul Puri, Nikolai Yakovenko, and Bryan Catanzaro. Practical text classification with large pre-trained language models. *arXiv preprint arXiv:1812.01207*, 2018.

[34] Wenhao Ying, Rong Xiang, and Qin Lu. Improving multi-label emotion classification by integrating both general and domain-specific knowledge. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 316–321, 2019.

[35] Nadir Durrani, Hassan Sajjad, and Fahim Dalvi. How transfer learning impacts linguistic knowledge in deep nlp models? *arXiv preprint arXiv:2105.15179*, 2021.

[36] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

[37] Federico Bianchi, Debora Nozza, and Dirk Hovy. Xlm-emo: Multilingual emotion prediction in social media text. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 195–203, 2022.

[38] Francesco Barbieri, Luis Espinosa-Anke, and Jose Camacho-Collados. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In *Proceedings of LREC*, 2022.

[39] Salar Mohtaj, Behnam Roshanfekr, Atefeh Zafarian, and Habibollah Asghari. Parsivar: A language processing toolkit for persian. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[40] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[41] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[42] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[43] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, June 2016. Association for Computational Linguistics.

[44] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[45] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[46] Colin Raffel and Daniel PW Ellis. Feed-forward networks with attention can solve some long-term memory problems. *arXiv preprint arXiv:1512.08756*, 2015.

[47] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.

[48] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*, 2016.