PSA43K: A Fine-Grained Persian Sentiment Analysis Dataset

Anonymous ACL submission

Abstract

Sentiment Analysis is an rapidly growing field of natural language processing (NLP). Persian language lacks human labeled datasets for sentiment analysis. Existing studies rely on small, low-quality datasets, limiting model accuracy and generalizability. To address this challenge, we present a Persian Sentiment Analysis dataset named PSA43K consisting of 43K real comments. Notably, this dataset was annotated by a team of native Persian speakers. Furthermore, we designed a model that achieved an accuracy rate of 64% in 5-class classification task, addressing the technical gap and advancing sentiment analysis for the Persian language.

1 Introduction

Sentiment analysis, a subfield of NLP, is an important task concerned with the determination of opinion and subjectivity in a text, which has many application [15, 13]. Massive user reviews available on e-commerce platforms are becoming valuable resources for both customers and merchants [4]. However, the accuracy of sentiment analysis models depends on the quality and size of the labeled training dataset [20].

Persian language lacks human labeled datasets for sentiment analysis. Existing studies rely on small, low-quality datasets, limiting model accuracy and generalizability. Shumaly et al. [18] conducted a study on Persian Sentiment Analysis using reviews from Digikala website. Ataei et al. [2] created a Persian dataset called Pars-ABSA, manually labeled from the Digikala website, and also employed deep learning methods for Sentiment Analysis. Pouromid et al. [16] labeled 12,000 Persian tweets for a dataset. Farahani et al. [12] proposed a monolingual model on Digikala and SnappFood comments, achieved notable results. Dashtipour et al. [7] developed a new framework that combines linguistic rules and deep learning to distinguish positive and negative sentiments. Basiri and Kabiri [3] suggested two new datasets, SPerSent and CNRC, and utilized majority voting and NB methods to identify the overall polarity of comments on Digikala website.

To address this issue, we've created a high-quality dataset for Persian Sentiment Analysis, consisting of 43,000 comments from Digikala(one of Iran's largest e-commerce platforms) and Snapfood (an online food ordering service). Each sentence is labeled with one of five class labels: positive, very positive, neutral, very negative, and negative. To ensure label accuracy, each sentence was annotated by at least three native Persian. The dataset contains over 43,000 annotated samples and professional annotators were hired and trained to identify different types of sentiments. Our work offers a valuable resource that enhances the development of sentiment analysis by offering a high-quality Persian dataset for training purposes. We make the following contributions:

- A novel sentiment analysis dataset is created which is considered the largest to date and that was annotated three times by human and went thought multiple rounds of quality checks to ensure high-quality annotations.
- A model is designed which achieves 64% of accuracy in 5-class classification task.
- The dataset will be made public following acceptance to the research community.

The paper is structured as follows: Section 2 reviews existing literature. Section 3 covers data collection, annotation, and analysis. Section 4 outlines the Model Development process, including architecture and hyperparameters. Section 5 provides performance evaluation details.

2 Related works

We begin with a survey of existing sentiment analysis datasets followed by the ML models.

2.1 Existing Datasets

DeepSentiPers [17] is a valuable dataset for NLP tasks related to digital product reviews. With a total of 12,138 user opinions, labeled with five different classes. A part of this dataset is obtained by data augmentation technique.

MirasOpinion [1] is valuable dataset for Persian NLP tasks, offering a large collection of user comments from Digikala. It was labeled using crowd-sourcing via a Telegram bot, resulting in 93,868 annotated documents with a balance of positive, negative, and neutral comments. The dataset is not publicly accessible.

Snapfood [14] is a leading online food delivery company that has curated a dataset of 70,000 user comments with just two polarity classifications: (0) Happy or (1) Sad.

Shumaly et al. [18] conducted a study on Persian Sentiment Analysis using reviews from the Digikala website. Three million Persian reviews were collected and pseudo-labeling technique is used to annotate the dataset and labels.

Dashtipour et al. [6] conducted a study on Persian Sentiment Analysis. They developed an 800-query dataset and created a context-aware framework that combined textual, acoustic, and visual features. The dataset has just 3 classes.

Ataei et al. [2] created a Persian dataset called Pars-ABSA, manually labeled from the Digikala website, and employed deep learning methods for Sentiment Analysis. This dataset has just 3 classes.

Basiri and Kabiri [3] suggested two new datasets, SPerSent and CNRC, and utilized majority voting and NB methods to identify the overall polarity of comments on the Digikala website. This datasets are suitable for binary classification.

2.2 Existing Machine Learning Models

Dashtipour et al. [5] applied Convolutional Neural Networks (CNN) and Long-Short-Term Memory (LSTM) algorithms, with LSTM demonstrating better performance. The study achieved impressive precision, recall, F1, and accuracy results up to 96%, 96%, 96%, and 95.61% respectively.

Dashtipour et al. [8] created an ensemble classifier for Persian Sentiment Analysis. By combining classic (SVM, MLP) and deep (CNN) machine learning classifiers with word2vec, they achieved an accuracy of 79.68%. The best performance was observed when using bigrams and combining the ensemble classifier with SVM, resulting in 80% precision, 79% recall, 75 F1 score, and 78.18% accuracy.

Zobeidi et al. [21] proposed a deep learning system for sentence-level review classification. It used matrices, CNN for feature extraction, and Bi-LSTM for review classification. On the Digikala dataset for mobile and cameras, they achieved high accuracy rates: 94% precision, 95% recall, 94% F1, and 95% accuracy.

Pouromid et al. [16] labeled 12,000 Persian tweets for a dataset, used to train ParsBERT, achieving 82% accuracy and outperforming a lexicon-based model.

Farahani et al. [12] proposed a monolingual model on Digikala and SnappFood comments. For Digikala, accuracy reached 82.52% and F1-value was 81.74%. SnappFood saw an accuracy of 87.8% and an F1-value of 88.12%.

Dashtipour et al. [7] developed a new framework that combines linguistic rules and deep learning to distinguish positive and negative sentiments. They evaluated two datasets: one with 3000 product reviews from Digikala and another with 3600 hotel reviews from Hellokish. For the Digikala dataset, they achieved 81.14% accuracy, 76% precision, 98% recall, and 84% F1 value. For the hotel dataset, they achieved 86.29% accuracy, 87% precision, 92% recall, and 89% F1 value.

Dehkharghani [10] introduced a new approach to detecting sentiment polarity in the Persian language by translating existing English language polarity lexicons into Persian. The author then used a supervised method like LR to assess the overall polarity of the translated words, achieving an accuracy rate of 95.92% and an F1 score of 96% using 5-fold cross-validation.

3 Methodology

In this section, we overview the methodology for creating the Persian Sentiment Analysis dataset, including data collection, annotation, and analysis.

3.1 Dataset

A big challenge in Sentiment Analysis for the Persian community is the lack of proper labeling tools such as Amazon Mechanical Turk. Due to various reasons, there is still a shortage of datasets necessary for AI research and NLP tasks in Persian. To solve this issue, our project aims to make a complete dataset for analyzing sentiments in Persian. We will explain the steps we took to ensure the

Figure 1: Telegram bot image

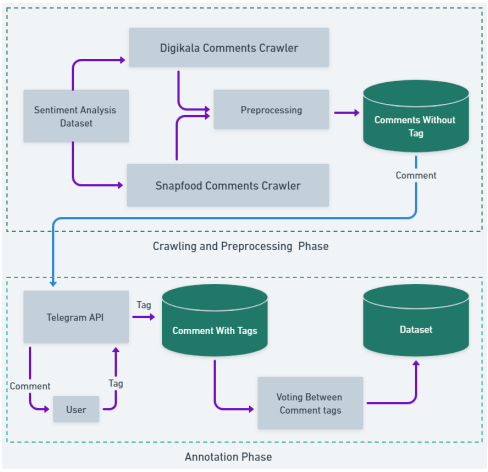

Figure 2: The key stages of dataset preparation

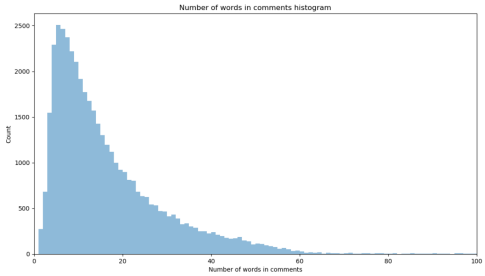

Figure 3: Comments length

dataset's quality and create a reliable resource for Persian Sentiment Analysis.

3.2 Gathering and Cleaning Data

We started a data collection process by scraping customer comments' from Digikala.

In the second step, we used Hazm [9], a Python package for Persian language processing, to normalize comments and cleaning the data. After checking the comments, we noticed most of them were positive which would result in an imbalanced dataset. To tackle this issue, we added some more comments from Snapfood. Since both applications are focused on selling online, it can be easy to mix the comments from the two platforms.

Our dataset had around 70,000 comments, but some issues affected the annotation of the dataset. Certain words, like ولی and اما (both meaning "but" in English), made it hard to classify comments into five classes. To ensure accuracy, we omitted less than 5% of comments with these words.

After additional data cleaning, we found comments using **"Fenglish"** where English letters are used to write Persian. To ensure data quality, we excluded such comments. The final dataset, free of duplicates, non-Persian comments, and Fenglish, consisted of 43,000 unique and clear Persian-language comments.

3.3 Labeling

In the next stage, we labeled comments using a **Telegram bot**. The bot sends the comments to taggers who would respond with the appropriate tags. To assist taggers and ensure consistency, the text and voice instructions for using the bot effectively were offered.

Figure 1 displays a Telegram bot with buttons designed for easily labeling comments.

To make sure tags are correct, three people tagged each comment, and the final tag was decided via a voting method. This method reduces errors and improves tag accuracy by soliciting opinions from different experts. Figure 2 shows the process of transforming raw data into an annotated dataset.

3.4 Dataset Analysis

To understand the properties of the dataset, we explore three aspects of the dataset: (i) **Comment length**, (ii) **Labels frequency**, and (iii) **normalized histogram of sentiment annotations at each different length comment**.

According to Figure 3, the majority of comments were between 5 and 25 words in length. This means users usually share their sentiment briefly, showing the importance of understanding emotions in shorter texts.

Figure 4 illustrates the label frequency of each class, providing insights into the distribution of sentiment annotations.

Figure 5 displays the distribution of the annota-

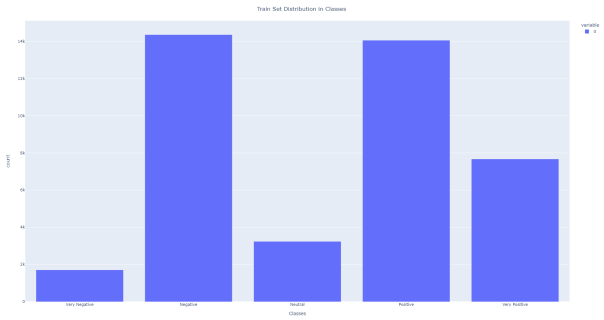

Figure 4: Labels frequency in each class

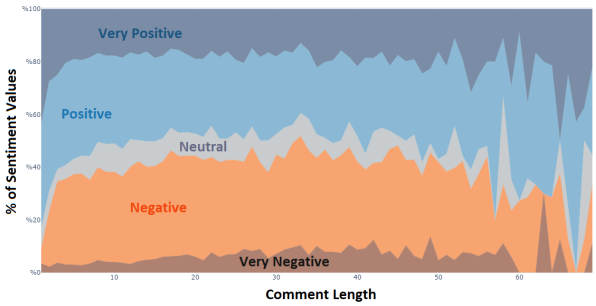

Figure 5: Distribution of sentiment annotations in the comments, based on their length

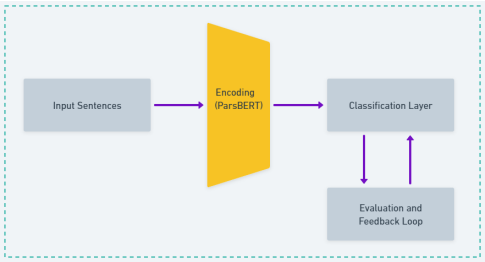

Figure 6: The architecture of our model

Table 1: Model Test Accuracy on PSA43K Dataset.

Model	5-class	3-class
ParsBERT	**64.0%**	**83.6%**
Multilingual BERT	59.9%	80.1%

tions in the comments, based on their length [19]. By normalizing the histogram and cropping it to the one to the seventy-word range, the plot highlights more details and provides a clear overview of the data. The distribution shows that the comments in the dataset have a well-balanced distribution.

4 Model Development

We used two transformer-based architectures: ParsBERT [14], a powerful pre-trained Persian model, and Multilingual BERT [11], a pre-trained model on a large corpus of multilingual data covering 104 various languages, to fine-tune the proposed model. During the training process, we used a batch size of 8 and trained the model for 4 epochs. We employed the Adam optimizer with a learning rate of 1e-5

Figure 6 shows our model architecture for sentiment analysis using PrasBERT and Multilingual BERT. Next, we evaluate the trained model's performance by measuring the standard evaluation metric of accuracy.

5 Evaluation

To evaluate the PSA43K dataset, we split it into three sets: 70% for training, 15% for validation, and 15% for testing. Also, the dataset was trained on two specific pre-trained language models, namely ParsBERT and Multilingual BERT.

We used two diffrent classification schemes: a 5-class classification (positive, very positive, neutral, very negative, and negative) and a 3-class (positive, neutral, and negative.) classification. Model performance was evaluated using the standard accuracy metric.

Table 1 presents a performance comparison of ParsBERT and Multilingual BERT models on the PSA43K dataset. ParsBERT achieved an accuracy of 64.0% in the 5-class and 83.6% in the 3-class, while Multilingual BERT achieved 59.9% accuracy in the 5-class and 80.1% in the 3-class.

ParsBERT outperformed Multilingual BERT in both 5-class and 3-class classifications, with the 3-class classification generally achieving higher accuracy for both models. Notably, ParsBERT showed a significant improvement in accuracy when moving from the 5-class to the 3-class.

6 Conclusion

In this paper, we present PSA43K, a novel Persian Dataset consisting of 43,000 comments for Sentiment Analysis; moreover, the method of collecting and annotating plus statistics of the dataset was discussed and demonstrated. At last, the corpus was evaluated with two models. Our dataset serves as a benchmark for developing new sentiment analysis models by researchers and practitioners. Future research should focus on other evaluation metrics and hyperparameter optimization to further improve our model's performance.

References

[1] Seyed Arad Ashrafi Asli, Behnam Sabeti, Zahra Majdabadi, Preni Golazizian, reza fahmi, and Omid Momenzadeh. 2020. Optimizing annotation effort using active learning strategies: A sentiment analysis case study in Persian. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2855–2861.

[2] Taha Shangipour Ataei, Kamyar Darvishi, Soroush Javdan, Behrouz Minaei-Bidgoli, and Sauleh Eetemadi. 2019. Pars-absa: an aspect-based sentiment analysis dataset for persian. *arXiv preprint arXiv:1908.01815*.

[3] Mohammad Ehsan Basiri and Arman Kabiri. 2017. Sentence-level sentiment analysis in persian. In *2017 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA)*, pages 84–89. IEEE.

[4] Jiahao Bu, Lei Ren, Shuang Zheng, Yang Yang, Jingang Wang, Fuzheng Zhang, and Wei Wu. 2021. ASAP: A Chinese review dataset towards aspect category sentiment analysis and rating prediction. In *North American Chapter of the Association for Computational Linguistics (NAACL) Conference: Human Language Technologies*.

[5] Kia Dashtipour, Mandar Gogate, Ahsan Adeel, Hadi Larijani, and Amir Hussain. 2021. Sentiment analysis of persian movie reviews using deep learning. *Entropy*, 23(5):596.

[6] Kia Dashtipour, Mandar Gogate, Erik Cambria, and Amir Hussain. 2021. A novel context-aware multimodal framework for persian sentiment analysis. *Neurocomputing*, 457:377–388.

[7] Kia Dashtipour, Mandar Gogate, Jingpeng Li, Fengling Jiang, Bin Kong, and Amir Hussain. 2020. A hybrid persian sentiment analysis framework: Integrating dependency grammar based rules and deep neural networks. *Neurocomputing*, 380:1–10.

[8] Kia Dashtipour, Cosimo Ieracitano, Francesco Carlo Morabito, Ali Raza, and Amir Hussain. 2021. An ensemble based classification approach for persian sentiment analysis. *Progresses in Artificial Intelligence and Neural Systems*, pages 207–215.

[9] Roshanak Davoodi and contributors. 2014. Hazm: A Python library for working with the Persian language. https://github.com/roshan-research/hazm.

[10] Rahim Dehkharghani. 2019. Sentifars: A persian polarity lexicon for sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(2):1–12.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

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

[13] Ehsan Hosseini-Asl, Wenhao Liu, and Caiming Xiong. 2022. A generative language model for few-shot aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 770–787.

[14] Marzieh Farahani Mohammad Manthouri Mehrdad Farahani, Mohammad Gharachorloo. 2020. Parsbert: Transformer-based model for persian language understanding. *ArXiv*, abs/2005.12515.

[15] Gaurangi Patil, Varsha Galande, Vedant Kekan, and Kalpana Dange. 2014. Sentiment analysis using support vector machine. *International Journal of Innovative Research in Computer and Communication Engineering*, 2(1):2607–2612.

[16] Mohammadjalal Pouromid, Arman Yekkehkhani, Mohammadreza Asghari Oskoei, and Amin Aminimehr. 2021. Parsbert post-training for sentiment analysis of tweets concerning stock market. In *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, pages 1–4. IEEE.

[17] Javad PourMostafa Roshan Sharami, Parsa Abbasi Sarabestani, and Seyed Abolghasem Mirroshandel. 2020. Deepsentipers: Novel deep learning models trained over proposed augmented persian sentiment corpus. *CoRR*, abs/2004.05328.

[18] Sajjad Shumaly, Mohsen Yazdinejad, and Yanhui Guo. 2021. Persian sentiment analysis of an online store independent of pre-processing using convol'utional neural network with fasttext embeddings. *PeerJ Computer Science*, 7:e422.

[19] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

[20] Yanyan Wang, Qun Chen, Jiquan Shen, Boyi Hou, Murtadha Ahmed, and Zhanhuai Li. 2021. Aspect-level sentiment analysis based on gradual machine learning. *Knowledge-Based Systems*, 212:106509.

[21] Shima Zobeidi, Marjan Naderan, and Seyyed Enayatallah Alavi. 2019. Opinion mining in persian language using a hybrid feature extraction approach based on convolutional neural network. *Multimedia Tools and Applications*, 78:32357–32378.