Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 7056–7060
Marseille, 20-25 June 2022

Pars-ABSA: a Manually Annotated Aspect-based Sentiment Analysis Benchmark on Farsi Product Reviews

Taha Shangipour Ataei[1], Kamyar Darvishi[1], Soroush Javdan[2], Behrouz Minaei-Bidgoli[1], Sauleh Eetemadi[1]
[1]Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
[2]School of Computer Science, Carleton University, Ottawa, Canada
{taha.pourataei, kamyar.darvishi}@gmail.com
soroushjavdan@cmail.carleton.ca
{b_minaei, sauleh}@iust.ac.ir

Abstract

Due to the increased availability of online reviews, sentiment analysis witnessed a thriving interest from researchers. Sentiment analysis is a computational treatment of sentiment used to extract and understand the opinions of authors. While many systems were built to predict the sentiment of a document or a sentence, many others provide the necessary detail on various aspects of the entity (i.e., aspect-based sentiment analysis). Most of the available data resources were tailored to English and the other popular European languages. Although Farsi is a language with more than 110 million speakers, to the best of our knowledge, there is a lack of proper public datasets on aspect-based sentiment analysis for Farsi. This paper provides a manually annotated Farsi dataset, Pars-ABSA, annotated and verified by three native Farsi speakers. The dataset consists of 5,114 POSITIVE, 3,061 NEGATIVE and 1,827 NEUTRAL data samples from 5,602 unique reviews. Moreover, as a baseline, this paper reports the performance of some aspect-based sentiment analysis methods focusing on transfer learning on Pars-ABSA.

Keywords: Sentiment Analysis, Aspect-based Sentiment Analysis, Opinion Mining, Farsi Language, Persian Language

1. Introduction and Background

Nowadays, by being in the era of data explosion where around 500 million tweets are sent daily, and since people are always curious about others' opinions, one challenge is to build a system to detect and summarize the overall sentiment of these data. Sentiment analysis is the computational study of detecting and extracting subjective information and attitudes about entities. The entity can represent individuals, events, products, or topics. The output of it is the opinion polarity. Polarity is generally expressed in different forms from two classes of **POSITIVE** and **NEGATIVE** or three classes of **POSITIVE**, **NEUTRAL**, and **NEGATIVE**, while at some researches, it is represented as a real number between 1-5 stars or 0-10 grade. Sentiment analysis was acknowledged in the early 2000s with(Turney, 2002), and (Pang et al., 2002), both of them doing binary classification on reviews. Sentiment analysis is generally performed at three different levels: document-based, sentence-based, and aspect-based. At both the document and sentence levels of sentiment analysis, the main goal is to detect the polarity of a specific document or a sentence. In contrast, aspect-based sentiment analysis is focused on identifying the polarity of the targets expressed in a sentence. A target is an object, its components, attributes and, features. For instance, at (Liu, 2010) a model is provided that identifies the polarity of product features that the reviewer has commented on. For example, in 'Food was great but the service was miserable.' There are two opinion targets, 'food' and 'service'. The reviewer has a **POSITIVE** sentiment polarity on 'food' and a **NEGATIVE** sentiment polarity on 'service'. This example shows why document-based and sentence-based systems are insufficient for this task. The superiority of aspect-based models to sentence-based and document-based models becomes vivid when manufacturers or service providers want to know which component or feature of their product is not well enough and needs improvement based on the negative reviews they get from customers. Generally, in aspect-based sentiment analysis, most of the data resources and systems built so far are tailored to English (Saeidi et al., 2016) and other languages like Chinese (Zhou et al., 2021; Bu et al., 2021) and Arabic (Al-Ayyoub et al., 2017; Al-Smadi et al., 2015). There are three datasets for English, which researchers mainly use to compare the performance of their systems which are Restaurants and Laptops (Pontiki et al., 2014) and Twitter (Dong et al., 2014). The first and second datasets contain annotated data samples from comments and reviews about laptops and restaurants from Semeval-2014 task 4: Aspect-based sentiment analysis. The last one is based on collected tweets from Twitter. Moreover at (Martens et al., 2021) authors gathered reviews from social media platforms like Twitter and Instagram on German language. Then, they manually annotated gathered data based on defined aspects in each review into **NEGATIVE**, **NEUTRAL** and **POSITIVE** categories. At last, they utilized BERT (Devlin et al., 2019) transformer model for classification. On the other hand, Farsi is the official language of Iran, Afghanistan, and Tajikistan and also is spoken in the east of Uzbekistan. Based on our knowledge, there are two datasets available for

this language. First, SentiPers (Hosseini et al., 2018) corpus that contains annotated data in all three levels (document-based, sentence-based, and aspect-based) with 21,375 target words and 26,996 corresponding opinion words identified from product reviews, which is highly imbalanced with more than 79% of them being labeled as **POSITIVE** and they claimed that have reached 63.15% score on polarity assignment in inner-annotator agreement. Second, ParsiNLU (Khashabi et al., 2021) is a Farsi benchmark for 6 various NLU tasks, which in aspect-based sentiment analysis they manually annotated 2,423 instances from reviews from two different domains of *food & beverages and movies* with 6 various labels from **VERY POSITIVE** to **VERY NEGATIVE** and **MIXED** but the class distribution reported on their paper shows that less than 5% of the annotated data was labeled as **NEUTRAL**.
The rest of the paper is organized as follows. In section 2, details about the data collection and annotation process are presented. In section 3 result of applying available systems for aspect-based sentiment analysis on the Pars-ABSA dataset is discussed. In section 4 we conclude and give future directions of research.

2. Dataset and annotation

This paper introduces a manually annotated aspect-based sentiment analysis corpus from customer reviews on products. It differs from past works mentioned earlier in various aspects, including the number of data instances, better inner-annotator agreement score, solving the imbalanced distribution of data, and providing reviews from different domains. Pars-ABSA dataset is available on a public repository [1].

2.1. Annotation

The data was gathered from the website of Digikala [2]. Digikala is the biggest e-commerce startup in Iran, and thousands of people buy goods from its website daily. Some of them submit comments about their purchased products and share their experiences with others. It is noteworthy to mention that more than 600,000 comments were scraped from the Digikala.Then, a framework based on python programming language was developed for manually annotating data instances. Furthermore, an annotation guideline was provided for annotators with a brief introduction to the task along with clearly expressed definitions of the classes and examples. Three native graduate students were employed to manually annotate the crawled data. It is important to note that all three annotators have annotated each data sample, and if two of them agree on the label, it was included in the dataset. In addition, to test the quality of their job and avoid any conflicts between annotators, after labeling the first 100 instances, a reviewer has been assigned to discuss the samples that they have disputes on them and fix misunderstandings.

Item	Value
# of targets	10,002
# of targets in train set	8,001
# of targets in test set	2,001
# of targets with positive polarity	5,114
# of targets with negative polarity	3,061
# of targets with neutral polarity	1,827
# of unique targets with positive polarity	1,494
# of unique targets with negative polarity	1,442
# of unique targets with neutral polarity	802
# of tokens	693,825
# of unique words	18,270
# of comments	5,602
Average # of words per comment	123.85
Average # of targets per comment	1.785

Table 1: Statistics of Pars-ABSA dataset.

2.2. Dataset Statistics

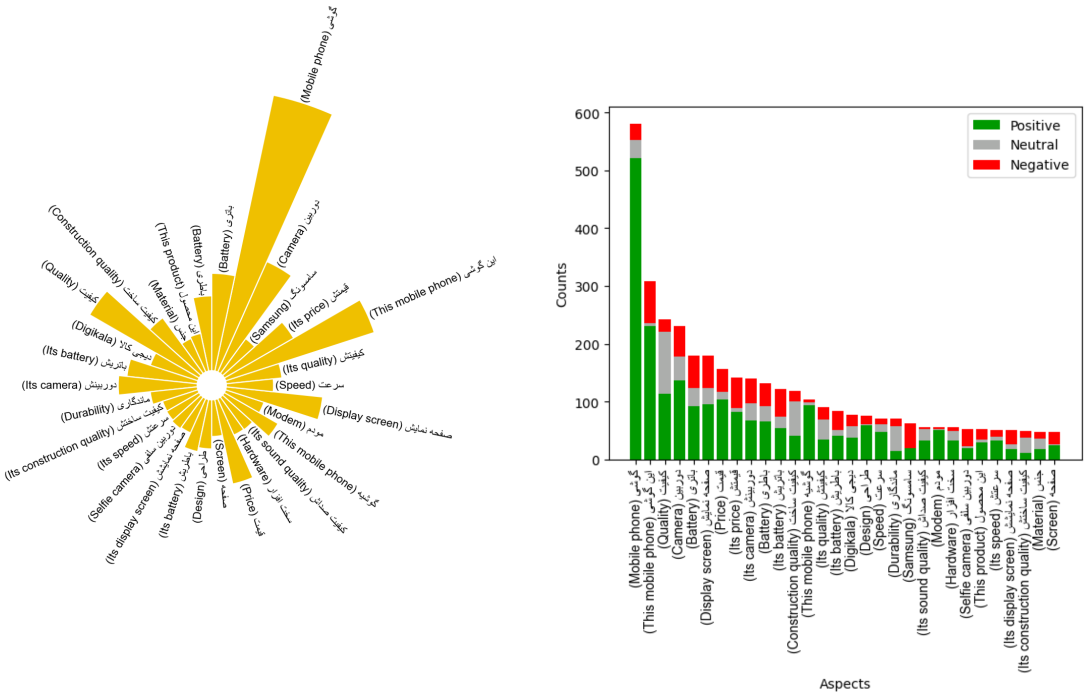
Statistical information about the proposed dataset is indicated in Table 1. Also, from 10,002 targets, the 30 most repetitive targets (e.g. گوشی *"Mobile phone"*, کیفیت*"Quality"* and دوربین *"Camera"*) is presented at Figure 1a additionally in Figure 1b for each target, the number of occurrences in each category is presented. For instance سامسونگ *"Samsung"* and دوربین سلفی *"Selfie Camera"* targets are mostly occurred in (**NEGATIVE**) category, expressing these two as the most unpleasant targets. As well, گوشی *"Mobile phone"* and طراحی *"Design"* are usually appeared in (**POSITIVE**) category that demonstrate them as the two most desirable targets. At last, کیفیت *"Quality"*, کیفیت ساخت *"Production Quality"* and ماندگاری*"Durability"* mostly took part in (**NEUTRAL**) category that explains reviewers can not decide on them to be good or bad.
Afterward, in Figure 2a frequency distribution of comments based on their lengths is presented and confirms that user reviews mostly have the length of less than 10 to 80 tokens. Moreover, in Figure 2b a density chart for comment lengths based on each category is given. It can be concluded from this chart that there is a relation between sentiment polarity of targets and the length of their reviews and when a review contains more than 800 tokens, the sentiment polarity of its target is usually labeled as (**NEUTRAL**).

2.3. Evaluation

To evaluate the quality of annotated corpus, it is common to calculate inter-annotator agreement. Because three annotators participated in this phase, Fleiss' (Fleiss, 1971) metric is computed as an inter-annotator agreement which is suitable for problems with more than two raters. In our case, we obtained 0.787 agreement overall annotated polarity of instances in the cor-

[1] https://github.com/Titowak/Pars-ABSA

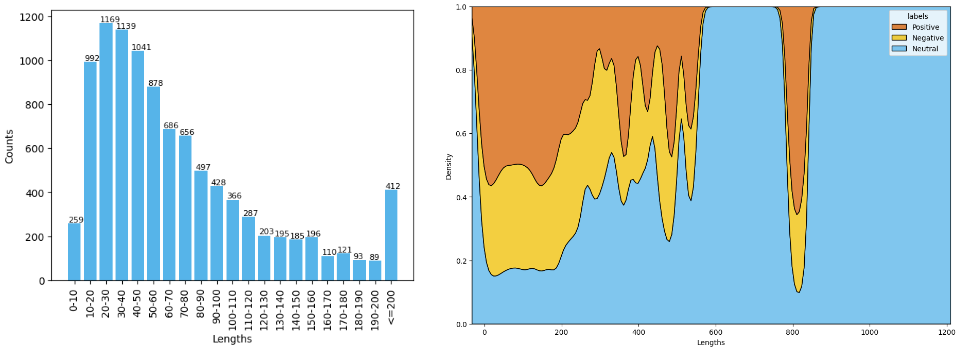
[2] `http://www.digikala.com`, Based on the terms of Digikala, the information of their website is allowed to be used for non-commercial activities with referring to them.



(a) Most repetitive targets

(b) Stacked representation of targets in each category

Figure 1: An overview of the 30 most repetitive targets in the Pars-ABSA dataset.



(a) Frequency distribution for comment lengths

(b) Density chart for comment lengths based on each category

Figure 2: An overview of comment lengths for each category and entirely

pus, that according to what is mentioned in (Fleiss, 1971) is considered as a substantial agreement.

2.4. Corpus structure

Pars-ABSA dataset is stored in two formats including XML and text. In XML format which is the main format of SemEval 2014 datasets, there is a main tag named ***sentences*** that contains all of the data instances. For each review in the dataset, there is a corresponding ***sentence*** tag available inside the main tag. ***sentence*** tag encompasses two types of tags, the first is ***text*** tag that contains the review and the second is ***aspectTerms*** that consists of one or more ***aspectTerm*** tags, as long as it is possible to have more than one aspect in each sentence. Each ***aspectTerm*** tag has four attributes, including the aspect, its polarity and, starting and ending index of it inside the review. An example of stored data instances in XML format is presented at Figure 3 in Appendix.

In the second format, for each aspect term, there are three corresponding lines inside the file, the review is at the first line, but the aspect term is replaced with ***T***, aspect term is written in the second line and in the third line, there is a number for sentiment polarity of the aspect term (1 for **POSITIVE**, 0 for **NEUTRAL** and -1 for **NEGATIVE**). An example of data instances in the text format is available in Table 2

3. Experiments

To evaluate Pars-ABSA corpus, it was split into two sets of training with 80% and test with 20% of data.

(English)	**Farsi**
In my opinion this speaker is in good shape and T is good too. its body material 1	در کل به نظر من اسپیکر خوش فرم و خوبی میاد T خوبی داره جنس بدنه ی ۱
I don't smell jasmine scent just pear scent. Although it's not a popular perfume I don't why its scent is so familiar for me and it feels that I have used it plenty of times. Generally its smell is good with a bad spreading and with medium T. durability 0	من اتفاقا بوی یاس احساس نمی کنم بیشتر بوی گلابی استشمام می کنم ولی با اینکه عطر متداولی نیست نمیدونم چرا بوش بنظرم تکراری اومد و انگار قبلا خیلی تجربش کرده بودم در کل خوش بو و با پخش بوی کم و T متوسط ماندگاری ۰
I don't smell jasmine scent just pear scent. Although it's not a popular perfume I don't why its scent is so familiar for me and it feels that I have used it plenty of times. Generally its smell is good with a bad T and with medium durability. spreading -1	من اتفاقا بوی یاس احساس نمی کنم بیشتر بوی گلابی استشمام می کنم ولی با اینکه عطر متداولی نیست نمیدونم چرا بوش بنظرم تکراری اومد و انگار قبلا خیلی تجربش کرده بودم در کل خوش بو و با T کم و ماندگاری متوسط پخش بوی -۱

Table 2: Samples from Pars-ABSA corpus in text format

Then, two systems based on transfer learning for aspect-based sentiment classification were used and trained on two specific pre-trained language models including, BERT (Devlin et al., 2019) multilingual base (cased) model, which is a language model trained on a large corpus of multilingual data on the top of 104 various languages and, ParsBERT (Farahani et al., 2021) which is based on BERT and was trained on large Farsi corpora of written materials that are publicly available. A brief description of mentioned methods is as follows:

- BERT (Devlin et al., 2019): This method is a general fine-tuning of the language models mentioned earlier with a linear classification at the last layer.

- LCF (Zeng et al., 2019): This method has performed significantly over the English datasets. It is based on multi-head self-attention and tries to focus on both the local and global contexts side by side. Therefore, it uses context features dynamic mask and context features dynamic weighted layers to recognize local context words and a BERT layer to catch long-term dependencies between local and global contexts.

Table 3 compares the performance of these systems with different language models on Pars-ABSA based on macro-average f1 score and accuracy metrics.

Analyzing the results achieved by models shows that as expected, LCF (Zeng et al., 2019) performs slightly better simple linear classification over the BERT(Devlin et al., 2019) since it employs an additional mechanism to focus on local context. Also, comparing pre-trained language models reveals that ParsBERT(Farahani et al., 2021) which is a monolingual model, outperforms the multilingual BERT(Devlin et al., 2019) model because it was explicitly pre-trained on a large amount of Farsi writing materials.

Model	**Multilingual**		**ParsBERT**	
	Acc	**F1**	**Acc**	**F1**
BERT	0.795	0.772	0.862	0.849
LCF	0.795	**0.78**	**0.874**	**0.863**

Table 3: Performance of models on Pars-ABSA corpus based on Accuracy and macro-average F1 metrics.

4. Conclusion and future works

In this paper, Pars-ABSA, a Farsi aspect-based sentiment analysis corpus was presented; moreover, the method of collecting and annotating plus statistics of the dataset was discussed and demonstrated. At last, the corpus was evaluated with models previously used for English datasets and, their performances were analyzed.

As future plans, our goal is to extend Pars-ABSA to include different domains such as restaurants and hotels and advanced pre-processing techniques since the reviews mostly have informal writing structures.

5. Bibliographical References

Al-Ayyoub, M., Gigieh, A., Al-Qwaqenah, A., Al-Kabi, M., Talafha, B., and Alsmadi, I. (2017). Aspect-based sentiment analysis of arabic laptop reviews. 12.

Al-Smadi, M., Qawasmeh, O., Talafha, B., and Quwaider, M. (2015). Human annotated arabic dataset of book reviews for aspect based sentiment analysis. In *2015 3rd International Conference on Future Internet of Things and Cloud*, pages 726–730.

Bu, J., Ren, L., Zheng, S., Yang, Y., Wang, J., Zhang, F., and Wu, W. (2021). ASAP: A Chinese review dataset towards aspect category sentiment analysis and rating prediction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2069–2079, Online, June. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., and Xu, K. (2014). Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54, Baltimore, Maryland, June. Association for Computational Linguistics.

Farahani, M., Gharachorloo, M., Farahani, M., and Manthouri, M. (2021). Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*, 53(6):3831–3847, Oct.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Hosseini, P., Ramaki, A. A., Maleki, H., Anvari, M., and Mirroshandel, S. A. (2018). Sentipers: A sentiment analysis corpus for persian.

Khashabi, D., Cohan, A., Shakeri, S., Hosseini, P., Pezeshkpour, P., Alikhani, M., Aminnaseri, M., Bitaab, M., Brahman, F., Ghazarian, S., Gheini, M., Kabiri, A., Mahabagdi, R. K., Memarrast, O., Mosallanezhad, A., Noury, E., Raji, S., Rasooli, M. S., Sadeghi, S., Azer, E. S., Samghabadi, N. S., Shafaei, M., Sheybani, S., Tazarv, A., and Yaghoobzadeh, Y. (2021). ParsiNLU: A Suite of Language Understanding Challenges for Persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162, 10.

Liu, B. (2010). Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition. Taylor and Francis Group, Boca*.

Martens, G., De Greve, L., Singh, P., Van Hee, C., Lefever, E., and Martens, G. (2021). Aspect-based Sentiment Analysis for German: Analyzing Talk of Literature" Surrounding Literary Prizes on Social Media. In *Computational Linguistics in The Netherlands (CLIN 31)*.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics, July.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August. Association for Computational Linguistics.

Saeidi, M., Bouchard, G., Liakata, M., and Riedel, S. (2016). SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1546–1556, Osaka, Japan, December. The COLING 2016 Organizing Committee.

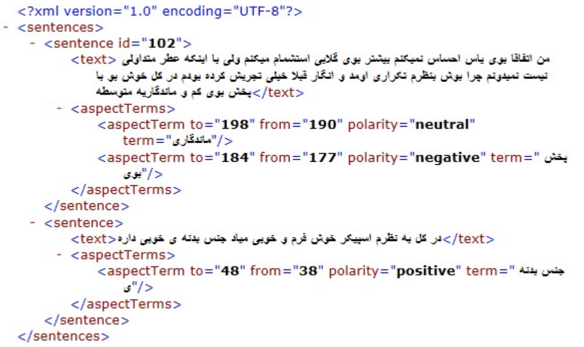
Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zeng, B., Yang, H., Xu, R., Zhou, W., and Han, X. (2019). Lcf: A local context focus mechanism for aspect-based sentiment classification. *Applied Sciences*, 9(16).

Zhou, J., Zhao, J., Huang, J. X., Hu, Q. V., and He, L. (2021). Masad: A large-scale dataset for multimodal aspect-based sentiment analysis. *Neurocomputing*, 455:47–58, Sep.

6. Language Resource References

A. Appendix



```
  <?xml version="1.0" encoding="UTF-8"?>
- <sentences>
  - <sentence id="102">
      <text> من اتفاقا بوی یاس احساس نمیکنم بیشتر بوی گلابی استشمام میکنم ولی با اینکه عطر متداولی
        نیست نمیدونم چرا بوش بنظرم تکراری اومد و انگار قبلا خیلی تجربش کرده بودم در کل خوش بو با
        پخش بوی کم و ماندگاریه متوسطه</text>
    - <aspectTerms>
        <aspectTerm to="198" from="190" polarity="neutral"
          term="ماندگاری"/>
        <aspectTerm to="184" from="177" polarity="negative" term=" پخش
          بوی"/>
      </aspectTerms>
    </sentence>
  - <sentence>
      <text>در کل به نظرم اسپیکر خوش فرم و خوبی میاد جنس بدنه ی خوبی داره</text>
    - <aspectTerms>
        <aspectTerm to="48" from="38" polarity="positive" term=" جنس بدنه
          ی"/>
      </aspectTerms>
    </sentence>
  </sentences>
```

Figure 3: An example of data samples stored in XML format