ParsFEVER: a Dataset for Farsi Fact Extraction and Verification

Majid Zarharan†, **Mahsa Ghaderan**†, **Amin Pourdabiri**†, **Zahra Sayedi**†
Behrouz Minaei-Bidgoli†, **Sauleh Eetemadi**†, and **Mohammad Taher Pilehvar**‡
†Iran University of Science and Technology, Tehran, Iran
‡Tehran Institute for Advanced Studies, Tehran, Iran
{majid_zarharan,m_ghaderan,amin_pourdabiri,s_sayedi}@
comp.iust.ac.ir, {b_minaei,sauleh}@iust.ac.ir, mp792@cam.ac.uk

Abstract

Training and evaluation of automatic fact extraction and verification techniques require large amounts of annotated data which might not be available for low-resource languages. This paper presents ParsFEVER: the first publicly available Farsi dataset for fact extraction and verification. We adopt the construction procedure of the standard English dataset for the task, i.e., FEVER, and improve it for the case of low-resource languages. Specifically, claims are extracted from sentences that are carefully selected to be more informative. The dataset comprises nearly 23K manually-annotated claims. Over 65% of the claims in ParsFEVER are many-hop (require evidence from multiple sources), making the dataset a challenging benchmark (only 13% of the claims in FEVER are many-hop). Also, despite having a smaller training set (around one-ninth of that in Fever), a model trained on ParsFEVER attains similar downstream performance, indicating the quality of the dataset. We release the dataset and the annotation guidelines at https://github.com/Zarharan/ParsFEVER.

1 Introduction

The spread of false information can lead to severe social and political problems (Wang, 2017). It would be extremely difficult to detect and track false information manually, given that the abundance of available technology has made it possible for these to be produced at scale and disseminated rapidly. Therefore, there has been a lot of interest in developing natural language technologies for fact-checking (Thorne and Vlachos, 2018). Unfortunately, similarly to many other fields of NLP that rely on manually curated datasets, fact-checking has remained restricted to a few high-resource languages for which large-scale annotated datasets are available.

In this paper, we present ParsFEVER, the first Farsi fact extraction and verification dataset. The dataset opens room for research in fact-checking and verification on low-resourced languages. ParsFEVER is constructed based on FEVER (Thorne et al., 2018), the most widely used dataset for fact-checking and fake news detection in English. We collected 22,906 claims by altering sentences extracted from introductory sections of 358 popular articles from Farsi Wikipedia. Annotators manually classified these claims into SUPPORTED, REFUTED, or NOTENOUGHINFO based on the provided reference pages. In addition, the annotators tagged those sentences which they used as evidence for this classification. Therefore, the dataset can be used for both fact-checking (a 3-class classification task) and evidence retrieval (which is a necessary step for the classification).

The quality of the dataset was evaluated using three different validation checks: (1) 5-way inter-annotator agreement, (2) agreement against super-annotators[1], and (3) manual validation by the authors. We also report experimental results for when ParsFEVER was used as a benchmark for the fact-checking task. In this task, given an input claim the model is expected to support or refute it and provide the corresponding evidence for this decision. If no enough evidence is found, NOTENOUGHINFO is returned. We evaluated the baseline system provided for FEVER on our dataset. The results indicate the more challenging nature of ParsFEVER: 50.0% (vs. 52.1% in FEVER) accuracy on a held-out test set on claim classification, and 28.1% (vs. 32.6% in FEVER) for evidence retrieval. Finally, we release ParsFEVER and related tools to allow further research on low-resource fact-checking, particularly in Farsi.

[1] The annotators who were responsible for training and leading other annotators.

Proceedings of the 10th Conference on Lexical and Computational Semantics, pages 99–104
August 5–6, 2021, Bangkok, Thailand (online) ©2021 Association for Computational Linguistics

2 Related Work

The only related datasets in Farsi are those of Zarharan et al. (2019) and Zamani et al. (2017). The former is a dataset for Farsi stance detection containing hundreds of instances in the news domain. Unlike ParsFEVER, the dataset does not provide any evidence for the claims; hence, it can only be used for a constrained fact-checking evaluation setting where evidences are already extracted for verifying stance. Also, the dataset of Zamani et al. (2017) is targeted towards rumor detection in Farsi tweets, which mostly relies on Twitter-specific features such as user profile information and response/retweet structure. In contrast, our dataset mostly focuses on lexical features.

ParsFEVER is mainly based on FEVER, a dataset widely used for fact extraction and verification in English. The dataset consists of around 185K claims generated by modifying sentences extracted from Wikipedia. The claims are classified as SUPPORTED, REFUTED, and NOTENOGHINFO. Despite being based on FEVER, our dataset has some fundamental differences that aim at making a more challenging benchmark for low-resourced languages. In the following section, we elaborate on the construction procedure of our dataset and the differences it has to that used for FEVER.

Other related datasets include HOVER (Jiang et al., 2020) and LIAR (Wang, 2017). HOVER is a dataset for many-hop fact extraction and claim verification. Unlike our dataset, which consists of single sentence claims, HOVER includes claims from one sentence up to one paragraph. It consists of 26K claims with SUPPORTED or NOTSUPPORTED labels. LIAR was instead derived from the short statements extracted from POLITIFACT.COM for fake news detection. This dataset contains 12.8K human-labeled instances.

Other related datasets in the social media domain include PHEME (Zubiaga et al., 2016b) and RumourEval (Zubiaga et al., 2016a). PHEME consists of 5,802 comment threads collected from Twitter, with approximately 103K tweets. This dataset has 1,972 and 3,830 threads labeled as rumour and non-rumour, respectively, resulting in an imbalanced dataset. RumourEval was released as part of the SemEval-2017 Task 8 (Derczynski et al., 2017). The dataset contains 330 rumour threads (4,842 tweets) from Twitter, annotated for both stance and veracity.

3 Dataset

Performing accurate fact-checking at scale requires a high-quality dataset along with the necessary algorithms and models. While there is a significant volume of research on the algorithms and models, they are generally language-agnostic. However, the datasets must be developed for each language independently. In this work, while using FEVER as a baseline, we modify their approach to make it more suitable for low-resource languages like Farsi.

Thorne et al. (2018) processed the June 2017 Wikipedia dump with Stanford CoreNLP (Manning et al., 2014) to collect sentences from the introductory sections of approximately 5K popular pages. In addition to this set of *primary* pages, all the related (*secondary*) pages[2] are retrieved. Following this procedure, we manually selected a set of 358 articles from the most popular Farsi pages crawled from fa.wikipedia.org. While FEVER provides an annotation tool, it leverages proprietary services which are not publicly available. Hence we developed our own Wikipedia crawler and annotation tools, which we release along with our dataset and annotation guidelines.

Table 1 shows two samples from ParsFEVER. In what follows in this section, we describe our procedure for constructing and validating the dataset.

3.1 Construction

The construction procedure of ParsFEVER consists of two phases; claim generation and claim labeling.

3.1.1 Phase 1 - claim generation

The objective of this phase was to generate claims for the 358 retrieved popular Wikipedia pages. We followed the following two steps.

(1) Sentence selection: In the construction of FEVER, this step was carried out in a random manner, i.e., a sentence was randomly selected from the corresponding Wikipedia page to serve as claim. Instead, we opted for a manual sentence selection. Specifically, each annotator was asked to carefully select a sentence from the introductory section of the corresponding page (primary page) in a way that directly relates to the article while containing as many (hyper-)links as possible. The last criteria were to guarantee a high number of many-hop claims. Many-hop[3] claims are essentially more

[2] Referenced pages from the main page.

[3] The number of hops of a claim is the same as the number of necessary evidence documents for the claim.

Verdict	*(English)*	Farsi
`supports`	**Claim:** Maryam Mirzakhani obtained the full score of the World Mathematical Olympiad in 1995 as an official student at the pre-university level. **Evidence:** *[Maryam_Mirzakhani]* In her junior and senior years of high school (Tehran Farzanegan School), she won a gold medal at the International Mathematical Olympiad in 1994 (Hong Kong) and 1995 (Canada). The following year, in Toronto, she became the first Iranian student to achieve a perfect score. *[Student]* A student is primarily a person who is under learning with the goal of acquiring knowledge. The term "student" denotes those enrolled in secondary schools and higher.	**Claim:** مریم میرزاخانی نمره کامل المپیاد جهانی ریاضی را در سال ۱۹۹۵ به عنوان محصل رسمی در سطح تحصیلات پیش از دانشگاه به دست آورد. **Evidence:** [مریم_میرزاخانی] میرزاخانی در دوران تحصیل در دبیرستان فرزانگان تهران، برنده مدال طلای المپیاد جهانی ریاضی در سال‌های ۱۹۹۴ (هنگ‌کنگ) و ۱۹۹۵ (کانادا) شد و در این سال به‌عنوان نخستین دانش‌آموز ایرانی نمره کامل را به دست آورد. [دانش‌آموز] دانش‌آموز از لحاظ لغوی به معنی کسی است که دانش می‌آموزد و در اصطلاح، برای اطلاق به محصلان رسمی در سطح تحصیلات پیش از دانشگاه به کار می‌رود.
`refutes`	**Claim:** Typhoid is not contagious at all. **Evidence:** *[Typhoid_fever]* Typhoid fever, also known as typhoid, is a disease caused by Salmonella serotype Typhi bacteria. *[Infection]* An infectious disease, also known as a transmissible disease or communicable disease, is an illness resulting from an infection. Some signs of infection affect the whole body, generally.	**Claim:** حصبه به هیچ وجه مسری نمی‌باشد. **Evidence:** [حِصْبه] حَصبه، تیفوئید یا تب تیفوئید یک بیماری عفونی است که در اثر عفونت باکتری Salmonella enterica سویه تیفی ایجاد می‌شود. [بیماری_عفونی] بیماری عفونی یا بیماری واگیر یا بیماری مسری (به انگلیسی: Infectious diseases یا transmissible diseases یا communicable diseases) به بیماری گویند که توسط عفونت منتقل و علائم و نشانه‌های بیماری ظاهر شود.

Table 1: Sample instances from ParsFEVER (English translations are shown for reference). For each instance, we show the claim, the corresponding label (verdict), and the evidence (text spans from Wikipedia articles, with the page title in brackets) used for this decision.

challenging as they require evidence retrieved from multiple pages. Specifically, we asked the annotators to produce their claims in a way that at least half of them would require information from other neighbouring Wikipedia pages (secondary pages, i.e., those pages that are linked within the original claim) with the help of a custom dictionary.[4] Consequently, more than 87% of the claims in FEVER need information from only a single Wikipedia page (one hop) (Jiang et al., 2020). However, over 65% of the claims in ParsFEVER are many-hop. After selecting an appropriate sentence, at least two and at most five claims were generated, constituting our set of original claims.

(2) Claim mutation: Following Thorne et al. (2018), we asked the annotators to mutate the original claims. Six types of mutations were considered: paraphrasing, negation, substituting an entity/relation with a similar/dissimilar one, and making the claim more general/specific. At most, five mutated claims were generated for each mutation type.

In both steps in claim generation, the annotators were asked to construct claims that only target one specific fact. This was to avoid multiple-target claims, which can potentially have contradictions. In addition, the claims are required to be based on the entity of focus on the primary page.

3.1.2 Phase 2 - claim labeling

In this stage, each mutated claim is labeled with one of the SUPPORTED, REFUTED, or NOTENOUGHINFO tags. This requires the annotators to identify the appropriate evidence. The annotator specifies one of the SUPPORTED and REFUTED tags only when a strong evidence exists: SUPPORTED if the reason supports the claim, and REFUTED otherwise. If this decision needs additional knowledge (dic-

[4] The dictionary comprises the list of terms (hyper)linked in the original sentence and all the other sentences from the corresponding Wikipedia page.

tionary), the evidence has to be updated with the corresponding new extra entries. Finally, in case the information on Wikipedia pages is not enough to justify the verdict, the claim is labeled as NOTENOUGHINFO.

To simplify the annotation process, we provide all sentences from the introductory section of the primary and secondary pages. We let the annotators use any combination of these sentences as evidence. In contrast, Thorne et al. (2018) just provided the first sentence of each secondary page. Thorne et al. (2018) defined the dictionary using the title of secondary pages and their first sentence. It is worth mentioning that the first sentence might not necessarily offer any valuable extra information. The annotators could easily add an arbitrary Wikipedia page by providing its URL. As a result, the system automatically adds all sentences from the introductory section of the page and its dictionary. At last, by using all the provided sentences in the annotation interface, the annotators record the sentences necessary to justify their verdict.

3.2 Annotators

Our annotation team had 14 native Farsi speakers, all of whom were involved in phase 1 and phase 2. All the annotators were trained for the task prior to the annotation. There was no intervention during the annotation process, and annotators were paired randomly for various instances in phase 2.

3.3 Validation

During claim labeling (task 2), we carried out a verification step to filter out noisy claims. As a result, around 2% of all generated claims were skipped by annotators for not satisfying the required quality criteria. Approximately 1% contained typos, and about 5% were flagged as too ambiguous, all of which were excluded from our dataset.

We implemented three forms of data validation for claim labeling: 5-way inter-annotator agreement, an agreement against super-annotators, and manual validation by the authors. To this end, we selected 3% of claims to be annotated by five annotators and calculated a 5-way inter-annotator agreement. The Fleiss k score was computed as 0.599, which is lower than that reported for FEVER (0.684). This can be attributed to the fact that ParsFEVER comprises significantly more many-hop instances, making the annotation task more challenging. Also, Table 2 shows the results of agreement against super-annotators of ParsFEVER compared to FEVER: 12 of the 14 annotators had an agreement of 87% with the super-annotators (the other two had 81% and 79%).

	Precision	**Recall**	**F1**
FEVER	95.42	72.36	82.30
ParsFEVER	86.95	85.23	86.08

Table 2: Agreement against super-annotators of ParsFEVER compared to FEVER.

	FEVER	**ParsFEVER**
IAA	0.84	0.71
Human	0.75	0.63

Table 3: The agreement of 500 randomly selected claims from ParsFEVER compared to FEVER (in terms of accuracy). IAA and Human respectively stand for Inter-annotator agreement and annotators' agreement against gold labels.

We also randomly selected 500 claims from ParsFEVER and FEVER to make another comparison. We asked two annotators to label each claim of the selected set for FEVER and ParsFEVER. Table 3 shows evidence and label agreement. The agreement of ParsFEVER is lower than FEVER. This is because most ParsFEVER claims are many-hop, resulting in a more challenging dataset (Jiang et al., 2020). Finally, if we ignore the correct evidence for ParsFEVER, the inter-annotator agreement and annotators' agreement against the dataset are 0.92 and 0.87 based on accuracy, respectively.

3.4 Dataset Statistics

Table 4 lists the distribution of instances across the three classes in the training, development, and test sets. Unlike FEVER which only includes mutated claims, in ParsFEVER we consider both mutated and original claims to improve training.

4 Experiments

Following Thorne et al. (2018), we implemented a full pipeline system for fact verification and extraction with the following three modules:

1. A document retrieval component (Chen et al., 2017) to find the most relevant page to a specific claim.

2. A sentence retrieval module to extract the evidence sentence (DrQA-based sentence retrieval module).

Split	SUP	REF	NEI
Training	6,253	4,008	5,685
Dev	841	824	861
Test	853	833	863
Total	7,947	5,665	7,409

Table 4: Distribution of instances in ParsFEVER across the three classes: SUPPORTED (SUP), REFUTED (REF), and NOTENOUGHINFO (NEI). The statistics are for the pruned dataset, i.e., after omitting claims which are ambiguous or contain typo (around 1,885 samples). Both mutated and original claims are included in the dataset.

3. Two recognizing textual entailment (RTE) models are used to classify the claim based on collected evidences as SUPPORTED, REFUTED, or NOTENOUGHINFO.

These models are based on MLP (Riedel et al., 2017), with a single hidden layer that benefits term frequencies and TF-IDF cosine similarity between the claim and evidence, and Decomposable Attention (Parikh et al., 2016, DA). Given that NOTENOUGHINFO instances are not associated with any evidence, they cannot be used for training the RTE models. To address this issue, Thorne et al. (2018) proposed two alternatives solutions: sampling a sentence (as evidence) from the nearest page to the claim (NP) or using the document retrieval component to uniformly select a random sentence (as evidence) from Wikipedia (RS).

4.1 Results

We customized the system based on Farsi. Following Thorne et al. (2018), we set $k = 5$ (k nearest documents to the claim for document retrieval) and $l = 5$ (top l-most similar sentences from the k-most relevant documents). We also checked for other values of the two parameters. However, no improvements were observed on the development set of ParsFEVER.

Table 5 shows the accuracy of the system on ParsFEVER and FEVER. ScoreEv and NoScoreEv respectively stand for accuracy score with respect to correct evidence retrieval and without considering the evidence. The first row in the table belongs to the best result reported by Thorne et al. (2018) using a decomposable attention model (DA) trained on NP. We show results on ParsFEVER using the full pipeline system when either NP or RS methods are used to provide evidence for NOTENOUGHINFO instances. DA generally performs better than MLP, particularly when combined with the NP strategy for sampling sentences. In fact, the best accuracy was achieved by DA/NP, with (ScoreEv) and without (NoScoreEv) the requirement to provide correct evidence with 28.06% and 50.02%, respectively.

Model		Accuracy (%)	
		NoScoreEv	ScoreEv
FEVER	DA/NP	52.09	32.57
ParsFEVER	MLP/NP	41.03	17.46
	MLP/RS	43.76	14.62
	DA/NP	50.02	28.06
	DA/RS	48.08	19.06

Table 5: Accuracy performance of FEVER's full pipeline system on ParsFEVER (best results for FEVER are reported).

5 Conclusion

We presented ParsFEVER, a novel and publicly available dataset for Farsi fact extraction and verification. We elaborated the construction procedure for this dataset, which focuses on having a rich dataset suitable for low-resource languages. Although this work uses Wikipedia as its source, other textual structures and corpora can also be used for fact extraction in this framework.

We evaluated the baseline system proposed for FEVER on our dataset. However, there have been recent developments in the field of fact-checking with models such as QABriefs (Angel et al., 2020). An immediate future work would be to take ParsFEVER as a more challenging benchmark (than FEVER) with significant many-hop operations as a benchmark for evaluating and analyzing existing fact-checking models. This analysis can also shed light on the ability of these models to go beyond the English languages and in low-resource settings.

6 Acknowledgments

The authors would like to thank the annotation team (especially Sepideh Sedghi, Masoume Afshari, Fatemeh Abvali, Baran Barbarestani and Najmeh Keynezhad) for their significant contribution in generating and labeling the claims. In addition, we would like to thank Mohammad Javad Pirhadi, who developed the annotation tools.

References

Jason Angel, Segun Taofeek Aroyehun, Antonio Tamayo, and Alexander Gelbukh. 2020. NLP-CIC at SemEval-2020 task 9: Analysing sentiment in code-switching language using a simple deep-learning classifier. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 957–962, Barcelona (online). International Committee for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the fake news challenge stance detection task. *CoRR*, abs/1707.03264.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

William Yang Wang. 2017. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Somayeh Zamani, Masoud Asadpour, and Dara Moazzami. 2017. Rumor detection for persian tweets. In *2017 Iranian Conference on Electrical Engineering (ICEE)*, pages 1532–1536.

Majid Zarharan, Samane Ahangar, Fatemeh Sadat Rezvaninejad, Mahdi Bidhendi, Mohammad Taher Pilehvar, Behrouz Minaei, and Sauleh Eetemadi. 2019. Persian stance classification data set. In *Proceedings of the conference for Truth and Trust Online*.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016a. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11(3):1–29.

Arkaitz Zubiaga, Geraldine Wong Sak Hoi, Maria Liakata, and Rob Procter. 2016b. PHEME dataset of rumours and non-rumours. *figshare, Jun*.