Artificial Intelligence Applications And Innovations Vol. 1 No. 2 (2024)

Article history:
Received 10 November 2023
Revised 3 January 2024
Accepted 12 January 2024
Published online 02 April 2024

naab: A ready-to-use plug-and-play corpus for Farsi

Sadra. Sabouri[1*], Elnaz. Rahmati[1], Soroush. Gooran[1], Hossein. Sameti[1]

[1] Speech and Language Processing Lab, Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

*** Corresponding author email address**: sadra@ee.sharif.edu

A r t i c l e I n f o

Article type:
Original Research

How to cite this article:
Sabouri, S., Rahmati, E., Gooran, S., & Sameti, H. (2024). naab: A ready-to-use plug-and-play corpus for Farsi. Artificial Intelligence Applications and Innovations, *1*(2), 1-8.
https://doi.org/10.61838/jaiai.1.2.1

A B S T R A C T

The rise of large language models (LLMs) has transformed numerous natural language processing (NLP) tasks, yet their performance in low and mid-resource languages, such as Farsi, still lags behind resource-rich languages like English. To address this gap, we introduce Naab, the largest publicly available, cleaned, and ready-to-use Farsi textual corpus. Naab consists of 130GB of data, comprising over 250 million paragraphs and 15 billion words. Named after the Farsi word ناب (meaning "pure" or "high-grade"), this corpus is openly accessible via Hugging Face, offering researchers a valuable resource for Farsi NLP tasks. In addition to naab, we provide naab-raw, an unprocessed version of the dataset, along with a pre-processing toolkit that allows users to clean their custom corpora. These resources empower NLP researchers and practitioners, particularly those focusing on low-resource languages, to improve the performance of LLMs in their respective domains and bridge the gap between resource-rich and resource-poor languages.

Keywords: Natural Language Processing, Low-resource Languages, Large Language Models, Textual Corpus, Open-source Dataset, Data Preprocessing, Persian Language Resources, Text Mining

1. Introduction

Large language models (LLMs) have revolutionized how people interact with technology, representing one of the most significant breakthroughs of the modern era [1, 2]. While these models have shown remarkable improvements across a wide range of tasks [3] in English, their performance in low and mid-resource languages, such as Farsi, often lags behind [4, 5].

Pre-training LLMs to produce Pretrained Language Models (PLMs) [6, 7] requires vast amounts of data, and large textual corpora are essential for fine-tuning these models for specific languages. Given this process's time- and resource intensive nature, having a readily available, large-scale corpus can significantly benefit researchers working to improve NLP in low-resource languages.

The lack of large-scale Farsi text data has made fine-tuning large language models challenging [8, 9]. This limitation often restricts the ability to train such models to only a handful of well-funded companies or countries, creating an uneven playing field. As a result, this lack of accessibility can hinder progress in open science, where collaboration and shared resources are essential for advancing NLP research in these underrepresented languages.

The largest previously available cleaned Farsi textual corpus was a 70GB dataset compiled from eight sources: Common Crawl - fa [10], Miras Text [11], W2C – Web to Corpus [12], Persian Wikipedia [13], Leipzig Corpora [14], VOA corpus [15], Persian poems corpus [16], and the Tehran English-Persian parallel corpus (TEP) [17]. This corpus has been cleaned and is available for direct download.

Meanwhile, several toolkits have been developed to streamline NLP workflows, including fine-tuning large models. These tools aim to democratize access to advanced NLP techniques and promote open science. A notable example is the Python library transformers [18], which has become the standard for training and fine-tuning transformer-based models. Hugging Face has also introduced a range of integrated libraries that enhance accessibility and collaboration in various NLP tasks.

Among these is the datasets library [19], which provides open-source corpora that are easily accessible to NLP researchers. However, none of the existing Farsi corpora are available on datasets. The first contribution of this work is to provide an easily accessible Farsi corpus, available to all through Hugging Face datasets.

One of the other primary challenges faced by NLP researchers is effective data pre-processing. Textual corpora, often derived from web-crawled data, frequently contain undesirable text and personal information. Traditional methods, such as trimming to remove unwanted patterns [6, 20], are often computationally expensive and memory-intensive. In this technical report, we introduce a more efficient alternative: a streaming pipeline for pre-processing texts in Farsi, which addresses these issues in a streamlined and resource-efficient manner.

Our solution to these challenges is embodied in the *naab* project, derived from the Farsi word ناب, meaning "pure" or "high-grade" [21]. The corpus provides 126GB of training data, consisting of more than 224 million sequences and nearly 15 billion words, and 2.3GB of test data, containing nearly 11 million sequences and 300 million words.

The main contributions of this project include the release of the largest cleaned and open-source Farsi corpus, naab, hosted on Hugging Face for easy accessibility; and the introduction of an easy-to-use, streaming-based pre-processing approach that enhances efficiency while maintaining high data quality.

Table 1

Statistics of the cleaned corpora included in the naab dataset. This table presents the size (in GB), the total number of paragraphs, the total number of words, and the average number per paragraph for each corpus.

Name	Size (GB)	#paragraphs	#words	#words/#paragraphs
Persian NLP	67	13,287,678	7,618,898,575	573.38
OSCAR-fa	36	60,099,393	4,193,005,807	69.76
AGP	23	141,912,688	2,776,681,752	19.56
LSCP	2.3	15,205,432	269,097,323	17.69
Telegram	0.9	6,471,586	100,253,032	15.49
Total	129.2	236,976,777	14,957,936,489	63.12

2. Materials and Methods

This section describes the materials used throughout the project and the methods to prepare the cleaned version of *naab*, now available as an open-source dataset. The following subsections detail the incorporated base corpora and the pre-processing techniques applied to ensure high-quality data.

2.1. Base Corpus

We utilized several corpora to build *naab*, combining various sources of Farsi text to create a comprehensive and diverse dataset. Table 1 provides statistics such as the size of each corpus, the number of paragraphs, and the word counts.

The largest underlying corpus is the Persian NLP corpus, accounting for 67GB and 7.6 billion words across approximately 13.3 million paragraphs, with an average paragraph length of 573.38 words. As described in [13], this corpus aggregates eight separate corpora: Common Crawl (65GB), MirasText (12GB), Web to Cor pus (1GB), Persian Wikipedia (787MB), Leipzig Corpora (424MB), VOA corpus (66MB), Persian Poems Corpus (61MB), Tehran English-Persian Parallel Corpus (33MB). We used a cleaned version of this dataset and further processed it with our proposed preprocessor (see Section 2.2).

The two other major base corpora were OSCAR-fa and AGP. While they con tribute significantly to the number of paragraphs (60 million and 141 million, respectively), they exhibit much shorter average paragraph lengths, especially

AGP, where the average is only 19.56 words per paragraph. This reflects a higher volume of shorter, more fragmented text, potentially sourced from informal or social media content.

The AGP corpus, originally a private resource from ASR Gooyesh Pardaz*, has been publicly released through this project. This corpus contains over 140 million paragraphs, totaling 23GB after cleaning. It is a diverse mix of formal and informal text gathered from websites and social media.

The OSCAR corpus (Open Super-large Crawled Aggregated coRpus) [22] is a multilingual dataset. We used the *unshuffled-deduplicated-fa* subset of OSCAR, resulting in 36GB of data after cleaning.

The Telegram corpus, small in size (0.9GB), features shorter paragraphs, aver aging 15.49 words per paragraph, further supporting the informal, conversational nature of this dataset. Being a popular messaging platform in Iran, Telegram served as a source of informal Farsi text. We curated a list of channels that span various topics such as news, entertainment, sports, and more. While smaller in size, this dataset is rich in colloquial Farsi and provides an up-to-date snapshot of daily conversational language.

Lastly, the Large Scale Colloquial Persian Language Understanding (LSCP) dataset [23], originally containing approximately 120 million sentences, offers a middle ground between formal and informal content. We focused on the Farsi portion of the dataset, resulting in 2.3GB of cleaned text comprising 15.2 million paragraphs, with an average paragraph length of 17.69 words.

Overall, the total size of 129.2GB and the diverse word-to-paragraph ratios suggest that *naab* offers a rich blend of formal and informal text types.

2.2. *Pre-process*

We developed a custom pre-processing script that utilizes efficient Linux kernel tools to clean the data with minimal memory overhead. Unlike traditional methods that load entire datasets into memory for pattern matching — leading to slower processing speeds — our approach pre-processes approximately 1GB of data per minute on an Intel(R) Xeon(R) CPU E5-2699 v3 @ 2.30GHz, offering exceptional speed.

In contrast to memory-intensive approaches that require $O(n)$ memory for large datasets, our method operates in a streaming fashion, reducing the memory requirement to $O(1)$ by processing the data in chunks. This allowed us to preprocess the massive 130GB dataset on a system with only 16GB of RAM.

The script is customizable and available on GitHub†. Below, we outline the main steps of the pre-processing pipeline:

2.2.1. *Filtering Non-Farsi Characters*

In the first step, we defined a filter that allows only "proper" words to pass the filter through, defined as words containing:

- All 32 characters of Farsi (ا to ی)
- Arabic characters which are ubiquitous in Farsi
- Symbolic characters (like '.', '?', '-', ',' and their Farsi version)
- Half-space in Farsi ('<200c>')

2.2.2. *Unifying Arabic/Farsi Characters*

Many texts use different shapes of Farsi characters interchangeably affecting the text analysis. To address this issue, less frequent character variants were replaced with their common alternatives. The substitution rules are listed in Table 2.

Table 2

Substitution list of characters to their alternative

* http://asr-gooyesh.com/en/

† https://github.com/Sharif-SLPL/t5-fa/tree/main/preprocess

To be substituted	Alternative
ي ا ىّ	ی
ة ا ۀ	ه
ك	ک
إ	ا
ڒ	ر
ۆ	و

2.2.3. White Spaces

After filtering, consequent spaces are normalized to one space (" ") to ensure consistency. In addition, any empty line resulting from previous cleaning steps is removed.

2.2.4. Removing Short Lines

In this final step, lines with fewer than 5 words (controlled by a parameter) are removed to reduce noise in the dataset. This ensures that the remaining text is rich in content and suitable for downstream tasks.

3. Results

We have released two versions of the *naab* corpus to facilitate various research and development needs. These versions differ in the level of pre-processing applied, giving users the flexibility to choose based on their requirements. Both versions are hosted on Hugging Face's *datasets* library for easy access.

3.1. naab-raw

The first version, *naab-raw*, represents the raw, unprocessed compilation of our newly gathered texts combined with existing Farsi corpora. This version is ideal for users who wish to handle data cleaning and pre-processing themselves, allowing full customization to fit specific tasks and experiments.

We chose Hugging Face's *datasets* library for distributing *naab-raw*, leveraging its ability to efficiently handle large datasets and link to external sources. Users can access the raw dataset under the repository SLPL/naab-raw*. Additionally, we provide a pre-processing script (see Section 2.2) that can be modified to meet different data cleaning and formatting needs.

3.2. naab

The second version, *naab*, is a cleaned, ready-to-use dataset that has been preprocessed to remove noise, non-textual elements, and other irrelevant content. This version is designed for plug-and-play usage, making it especially convenient for users who want to directly get started on model training or fine-tuning without needing to perform additional data preparation.

With *naab*, users can take advantage of Hugging Face's selective download feature, which allows for downloading specific parts of the dataset rather than the entire corpus. This feature is particularly useful for those working with limited storage or those focusing on a specific subset of the corpus. The dataset is hosted under SLPL/naab†, and users can find further information, including detailed usage instructions, in the dataset card. This version is ideal for researchers and developers seeking a clean, structured corpus for training natural language models, sentiment analysis, or other linguistic tasks in Farsi.

Both versions of the corpus are continuously maintained and updated, ensuring that users always have access to the latest resources. Additionally, community contributions are encouraged, further enriching the corpus and its potential applications.

4. Experiments

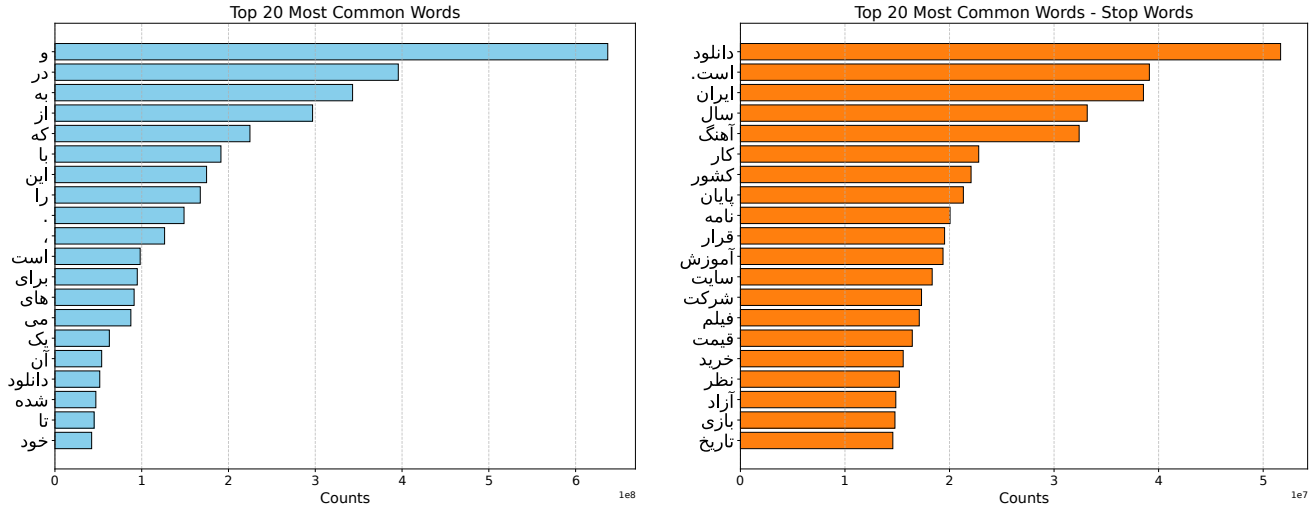
We analyzed the frequency distribution of words in the *naab* corpus. A comprehensive word count was performed,

* https://huggingface.co/datasets/SLPL/naab-raw

† https://huggingface.co/datasets/SLPL/naab

which provides the distribution of the most frequent terms. This analysis was conducted twice: once including all words in the corpus and once after removing common Farsi stop words [24] to provide a clearer view of the meaningful vocabulary (Figure 1).

The top 20 most frequent words, including stop words, are visualized in Figure 1(a). Expectedly, stop words like "و" (and) and "در" (in) dominate the list due to their high frequency in Farsi. However, after removing stop words, the most common words shift towards more meaningful terms, as shown in Figure 1(b). Once stop words were removed, frequently occurring terms shifted to content-bearing words. Notably, words such as "ایران" (Iran) and "دانلود" (Download) frequently appeared in the stop word-free list, reflecting the corpus's emphasis on regional topics influenced by internet-sourced content.



(a) (b)

Figure 1

The top 20 most common words in naab, along with their corresponding frequencies, are presented in two categories: a) including all words, and b) excluding stop words. The words are ranked by their counts, with the horizontal axis showing the frequency and the vertical axis listing the words.

5. Usage & Future Works

Naab corpus serves as an essential resource with broad applications across various domains of Natural Language Processing (NLP) and beyond. Given its size, diver sity, and richness, it offers numerous opportunities for both academic research and practical applications.

One of the primary uses of this corpus is for training large language models (LLMs). Self-supervised learning approaches, which leverage vast amounts of unlabeled text data, can use this corpus to develop powerful language models. Traditional models, such as n-gram language models, can use this dataset to capture word sequence patterns and short-term dependencies, providing useful insights into the structure of Farsi. More advanced transformer-based models, such as BERT [7], and BART [25], T5 [6], GPT [26], Llama [27], can also be pre-trained or fine-tuned on this corpus. With its vast and varied text, these models can significantly improve the understanding and generation of Farsi, making them ideal for a wide range of downstream tasks.

Researchers can use it to develop and improve text classification systems that categorize documents based on topics, sentiment, or other features. Named entity recognition (NER) models trained on this dataset will be able to detect entities such as names, locations, and dates from Farsi text. Additionally, part-of-speech (POS) tagging, which assigns grammatical categories to words [28], can be improved with this data. Text summarization systems can leverage this corpus to generate summaries of lengthy Farsi documents, making it a useful tool for information extraction and content consumption.

Beyond NLP tasks, this corpus contributes to advancements in speech process ing. Automatic Speech Recognition (ASR) systems, which convert spoken language into text [29, 30], can be trained and fine-tuned using this large collection of Farsi text, improving transcription accuracy. Text-to-Speech (TTS) models, which synthesize spoken language from text [30], will also use the corpus, producing more natural and fluent spoken Farsi.

These technologies are crucial for developing voice-based applications such as virtual assistants and automated customer service in Farsi.

The diversity of the corpus also makes it ideal for various types of linguistic research. Lexical and semantic studies will benefit from the large-scale dataset, as it provides rich material for exploring vocabulary evolution and usage patterns in Farsi.

Overall, this corpus is not only a critical resource for developing advanced NLP tools in Farsi but also a gateway to exploring the linguistic and cultural richness of the Farsi language.

6. Conclusions

The availability of large-scale textual data is a critical challenge for Farsi language researchers. In response to this, we present the largest open-source Farsi textual corpus, provided in both a cleaned version, referred to as *naab*, and a raw version, *naab-raw*. These two datasets are accessible to the research community as open-source resources hosted on Hugging Face's data hub, making them readily available for a wide range of NLP and linguistic studies. Furthermore, we introduce a stream-based pre-processing approach, enabling users to efficiently generate their own large-scale text datasets from scratch, accommodating those with specialized processing needs. This work aims to bridge a significant gap in Farsi language resources, empowering researchers and practitioners to push forward innovations in language technology.

7. Limitations

7.1. Stop words Retained for Contextual Integrity

In this work, we chose not to remove stop words to keep the meaningful structure of the text. Stop words, while often considered uninformative [31], can play a significant role in understanding the contextual relationships between words in natural language. However, the inclusion of these high-frequency words may affect certain statistical measures, such as n-gram frequency analysis, and could introduce noise in specific use cases where stop words are less informative.

7.2. Duplication Issues

Due to computational limitations, deduplication was not performed on this corpus. Given the large size of the text corpus, there is a possibility of duplicated content, as some of the underlying datasets may share the same texts. This duplication could influence the analysis, and affect the reliability of word embeddings. Future work could address this by implementing more advanced deduplication strategies to ensure more accurate and unbiased results. Researchers who wish to use a subset of this dataset can easily perform deduplication, provided they have sufficient computational resources. With manageable dataset sizes, deduplication processes become more feasible.

7.3. Personal Information

While we made significant efforts to ensure that the data sources used in the *naab* corpus avoid personal information, the nature of publicly available text data means there is still a potential for personal details to be present. To minimize this risk, we applied filtering techniques, including removing numerical data that could represent sensitive information like phone numbers, addresses, social security number, and credit cards. Despite these precautions, some identifiable data may still exist. Researchers and practitioners using this dataset should be cautious and adhere to ethical guidelines, especially if sensitive information is detected. It is the responsibility of users to ensure their work complies with all applicable legal, ethical, and institutional standards. This work and its authors do not take responsibility for any misuse of the dataset, and users are solely responsible for ensuring their usage follows appropriate data privacy protocols.

7.4. Corrupted Sentences

Since our pre-processing scripts filtered out specific characters, there is a slight possibility of introducing corrupted sentences in cases where some parts of the input text utilize a different character set or keyboard layout than others. Although this scenario is highly improbable, it remains a potential issue. To mitigate this risk, we analyzed part of the texts randomly and found no instances of such corruption. Nevertheless, we report this as a possible

limitation to acknowledge the theoretical risk and ensure transparency.

Authors' Contributions

The first author led the project, in charge of the primary responsibilities for data gathering, preprocessing, team coordination, experiment design, running analyses, and writing the technical report. The second author supported the first author in data collection, provided intellectual consultation, reviewed and double-checked the code, contributed to experiment design, and assisted in writing the paper. The third author contributed through technical consultation, negotiations for data gathering, experiment design, and assisting with paper writing. The fourth author served as the project advisor, providing guidance on the overarching goals of the project and performing a thorough review of the paper.

Declaration

The authors declare that all resources used in this work, including datasets and preprocessing scripts, are openly accessible to promote transparency and reproducibility. Links to these resources are provided throughout the paper. Researchers are encouraged to utilize, adapt, and contribute to these resources while adhering to ethical guidelines.

Transparency Statement

All data and code used in this study are openly available online to promote transparency and enable reproducibility. The *naab* and *naab-raw* datasets can be accessed through the Hugging Face datasets hub at SLPL/naab and SLPL/naab-raw, respectively. The preprocessing scripts, designed for efficient and customizable data cleaning, are hosted on GitHub. Researchers and practitioners are welcome to explore and adapt these resources for their needs.

Acknowledgments

We would like to thank everyone who has worked hard to promote open science and make resources accessible to all. A special thanks to Mohammadreza Hosseinian, CEO of ASR Gooyesh Pardaz, for kindly allowing us to use their private data and release it as open-source. We thank Sepand Haghighi who provided creative solutions to the challenges we faced and Mehran Ziadloo for his helpful feedback on the pre-print version of this project.

Declaration of Interest

The authors declare that they have no conflict of interest. The authors also declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This project received no external funding. All work was conducted independently without financial support from any organization or institution.

Ethical Considerations

The development of the corpus emphasized ethical data usage, with efforts to exclude sensitive or personal information. Filtering techniques were applied to remove numerical data, to minimize risks. However, due to the nature of publicly available text, some personal details may still exist. Researchers are urged to follow ethical guidelines and legal standards when using the dataset. The authors are not responsible for misuse, and users are solely accountable for adhering to data privacy and ethical practices.

References

[1] A. Srivastava *et al.*, "Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models," *arXiv preprint*, vol. arXiv:2206.04615, 2022. [Online]. Available: https://arxiv.org/abs/2206.04615.

[2] T. Teubner, C. M. Flath, C. Weinhardt, W. van der Aalst, and O. Hinz, "Welcome to the Era of ChatGPT et al. The Prospects of Large Language Models," *Business & Information Systems Engineering*, vol. 65, no. 2, pp. 95-101, 2023, doi: 10.1007/s12599-023-00795-x.

[3] Y. Chang *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1-45, 2024, doi: 10.1145/3641289.

[4] H. Avetisyan and D. Broneske, "Large language models and low resource languages: An examination of Armenian NLP," in *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, 2023, pp. 199-210, doi: 10.18653/v1/2023.findings-ijcnlp.18.

[5] S. Shen, L. Logeswaran, M. Lee, H. Lee, S. Poria, and R. Mihalcea, "Understanding the Capabilities and Limitations of Large Language Models for Cultural Commonsense," *arXiv preprint*, vol. arXiv:2405.04655, 2024, doi: 10.18653/v1/2024.naacl-long.316.

[6] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1-67, 2020. [Online]. Available: https://jmlr.org/papers/v21/20-074.html.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language

understanding," *arXiv preprint*, vol. arXiv:1810.04805, 2018. [Online]. Available: https://arxiv.org/abs/1810.04805.
[8] M. K. Habib, "The challenges of Persian user-generated textual content: A machine learning-based approach," *arXiv preprint*, vol. arXiv:2101.08087, 2021. [Online]. Available: https://arxiv.org/abs/2101.08087.
[9] S. Moniri, T. Schlosser, and D. Kowerko, "Investigating the Challenges and Opportunities in Persian Language Information Retrieval Through Standardized Data Collections and Deep Learning," *Computers*, vol. 13, no. 8, p. 212, 2024, doi: 10.3390/computers13080212.
[10] "Common crawl - farsi." https://commoncrawl.org/ (accessed.
[11] "MirasText." https://github.com/miras-tech/MirasText (accessed.
[12] M. Majliš, "W2C-Web to Corpus-Corpora," in *Proceedings of Corpus Linguistic*, 2011. [Online]. Available: https://scholar.google.com/citations?user=KrLDq0oAAAAJ&hl=nld. [Online]. Available: https://scholar.google.com/citations?user=KrLDq0oAAAAJ&hl=nld
[13] "Persian-Raw-Text." https://github.com/persiannlp/persian-raw-text (accessed.
[14] C. Biemann, G. Heyer, U. Quasthoff, and M. Richter, "The Leipzig corpora collection-monolingual corpora of standard size," in *Proceedings of Corpus Linguistic*, 2007. [Online]. Available: https://www.academia.edu/13038315/The_Leipzig_Corpora_Collection_Monolingual_corpora_of_standard_size. [Online]. Available: https://www.academia.edu/13038315/The_Leipzig_Corpora_Collection_Monolingual_corpora_of_standard_size
[15] J. Dehdari. "VOA news." https://jon.dehdari.org/corpora/ (accessed.
[16] A. Ghaderi. "Persian poems corpus." https://github.com/amnghd/Persian_poems_corpus (accessed.
[17] J. Tiedemann, "Parallel Data, Tools and Interfaces in OPUS," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, N. Calzolari *et al.*, Eds., 2012: European Language Resources Association (ELRA). [Online]. Available: https://aclanthology.org/L12-1246/. [Online]. Available: https://aclanthology.org/L12-1246/
[18] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38-45, doi: 10.18653/v1/2020.emnlp-demos.6.
[19] Q. Lhoest *et al.*, "Datasets: A Community Library for Natural Language Processing," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2021, pp. 175-184, doi: 10.18653/v1/2021.emnlp-demo.21.
[20] M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri, "ParsBERT: Transformer-based model for Persian language understanding," *Neural Processing Letters*, vol. 53, no. 6, pp. 3831-3847, 2021, doi: 10.1007/s11063-021-10528-4.
[21] Abadis, "Naab meaning," 2022. [Online]. Available: https://abadis.ir/fatofa/%D9%86%D8%A7%D8%A8/.
[22] J. Abadji, P. Ortiz Suarez, L. Romary, and B. Sagot, "Towards a cleaner document-oriented multilingual crawled corpus," *arXiv preprint*, vol. arXiv:2201.06642, 2022. [Online]. Available: https://arxiv.org/html/2410.23825v1.
[23] H. Abdi Khojasteh, E. Ansari, and M. Bohlouli, "LSCP: Enhanced Large Scale Colloquial Persian Language Understanding," *arXiv preprint*, vol. arXiv:2003.06499, 2020. [Online]. Available: https://www.researchgate.net/publication/373046387_Developing_an_Informal-Formal_Persian_Corpus.
[24] V. Kharazi and P. Kamalipour. "Persian (Farsi) Stop Words List - Persian." https://github.com/kharazi/persian-stopwords (accessed.
[25] M. Lewis, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," *arXiv preprint*, vol. arXiv:1910.13461, 2019, doi: 10.18653/v1/2020.acl-main.703.
[26] J. Achiam *et al.*, "GPT-4 technical report," *arXiv preprint*, vol. arXiv:2303.08774, 2023. [Online]. Available: https://scholar.google.nl/citations?user=HXUT9ZkAAAAJ&hl=th.
[27] A. Dubey *et al.*, "The LLAMA 3 herd of models," *arXiv preprint*, vol. arXiv:2407.21783, 2024. [Online]. Available: https://arxiv.org/abs/2407.21783.
[28] A. R. Martinez, "Part-of-Speech Tagging," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 1, pp. 107-113, 2012, doi: 10.1002/wics.195.
[29] H. Hadian, S. Gooran, S. Sabouri, S. Sadeghi, Y. Amini, and H. Sameti, "A review of the recent speech recognition methods," *Journal of Vibration and Sound*, vol. 11, no. 22, pp. 125-154, 2023. [Online]. Available: https://bibbase.org/show?bib=https%3A%2F%2Fbibbase.org%2Fnetwork%2Ffiles%2F6XPETtqWDEnoiQoeC&noBootstrap=1.
[30] C. L. Kao *et al.*, "Rapid development of an English/Farsi speech-to-speech translation system," in *Proceedings of the 5th International Workshop on Spoken Language Translation: Papers*, 2008, pp. 166-173. [Online]. Available: https://aclanthology.org/2008.iwslt-papers.4. [Online]. Available: https://aclanthology.org/2008.iwslt-papers.4
[31] D. J. Ladani and N. P. Desai, "Stopword Identification and Removal Techniques on TC and IR Applications: A Survey," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020: IEEE, pp. 466-472, doi: 10.1109/ICACCS48705.2020.9074166.