# PerPaDa: A Persian Paraphrase Dataset Based on Implicit Crowdsourcing Data Collection

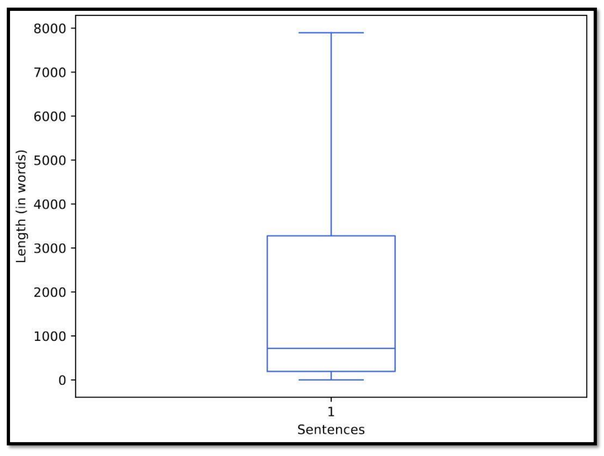
Salar Mohtaj Fatemeh Tavakkoli Habibollah Asghari

## 1. Introduction

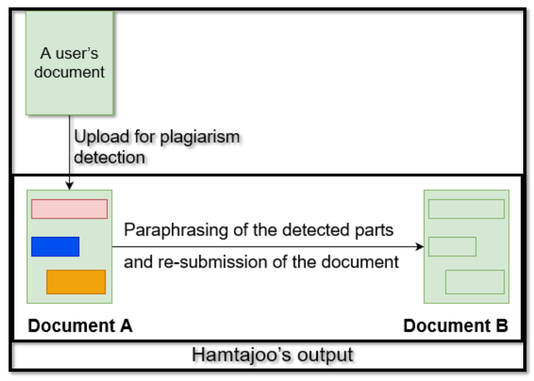
In this paper, a Persian paraphrase dataset is collected implicitly from a plagiarism detection system—Hamtajoo.
**PARAPHRASE DATASET** is a parallel corpus containing the original and the equivalent paraphrased pieces of text.
**IIMPLICIT CROWD-SOURCING** includes approaches where users do not necessarily know they are contributing.
**PERSIAN OR FARSI** is generally classified as western Iranian languages and is from the Indo-European family.
**HAMTAJOO** is a Persian plagiarism detection tool that is being used by journals, conferences, faculty members and students to detect cases of inadvertent or intentional text re-use in scientific papers.

## 2. Data Collection

- We have focused on those use cases in which users employ the plagiarism detection system to find case of text re-use and then try to conceal them by paraphrasing.
- Since the idea is to compare multiple submission of a document to extract those parts which paraphrased by users, we excluded the users who submitted just one document in the system.
- **18111** documents
- The length distribution



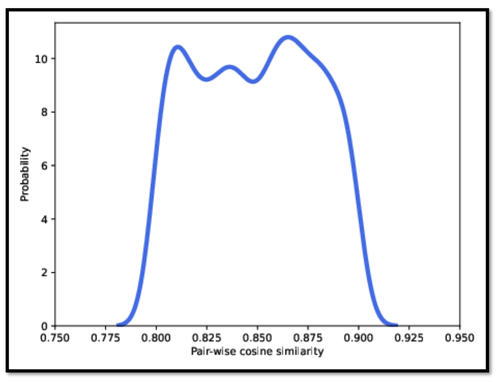

## 3. Corpus Construction

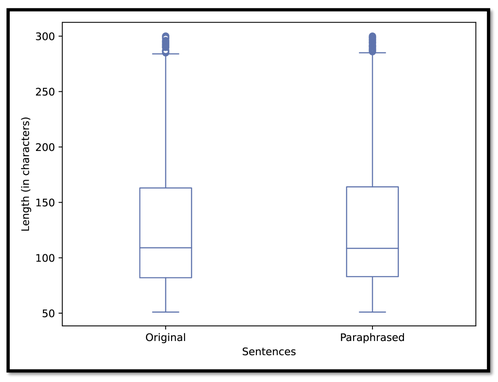



1. Detection of near duplicate documents for each user
   - cosine similarity between TF-IDF vectors of pairs of documents. We set the similarity threshold to [0.9 - 1)
2. Ordering documents in the near duplicate clusters based on the time of submission
   - The near duplicate documents are ordered based their submission date/time
3. Extracting sentences that detected as text matching cases in the lead documents by Hamtajoo
   - Tokenizing the whole text into sentences
4. Searching for the paraphrased sentences at the similar position in the subsequent documents
   - choosing the approximate location of the original sentence (in the lead document) in the subsequent documents.
   - The original and the potential paraphrased sentences have been embedded into vectors, using ParsBERT pre-trained model
   - The cosine similarity between the original sentence and the potential paraphrased sentences (choosing sentences that have at least 0.8 cosine similarity)
5. Post-processing of the extracted pairs
   - Applying some heuristics to exclude low quality pairs (short sentences, uncomplete sentences, and non-Persian sentences)

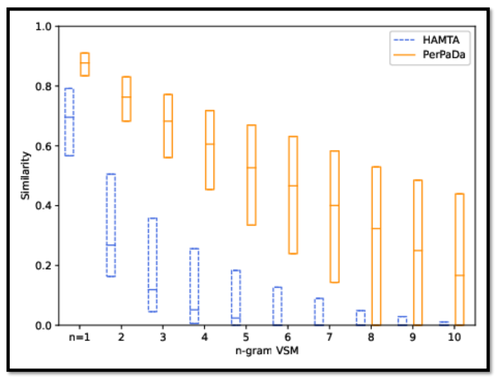
## 4. Evaluation

Dataset Statistics




Validation Result




German Research Center for Artificial Intelligence
Freie Universität Berlin
salar.mohtaj@tu-berlin.de
f.tavakkoli@fu-berlin.de
habib.asghari@ictrc.ac.ir