Datasheet for Azerbaijani Folktales Dataset

Nijat Jafarov, Namig Planov

January 2026

1 Datasheets for datasets

In this work, we adopt the Datasheets for Datasets framework proposed by Gebru et al. (2020) to document an Azerbaijani-language text corpus composed of folktales The dataset is collected from Azerbaijani Wikisource, which hosts texts that are either in the public domain or available under the Creative Commons Attribution-ShareAlike (CC BY-SA) license.

Given the scarcity of well-documented Azerbaijani NLP datasets, especially for low-resource language research, this datasheet aims to clearly describe the motivation, composition, collection process, licensing, and intended uses of the dataset. This documentation is intended to support transparency, reproducibility, and responsible reuse of the dataset in academic research.

2 Template

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for academic research and university-level projects in the area of Natural Language Processing (NLP). The primary motivation was to support experimentation, analysis, and model development for Azerbaijani.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

This dataset was created by two graduate students, Nijat Jafarov and Namig Planov enrolled in an NLP course at ADA University and The George Washington University. The dataset was developed as part of a university course project and was not created on behalf of any commercial organization, company, or external institution.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The creation of this dataset was not supported by any external funding, grant, or sponsorship. The dataset was developed independently as part of an academic course project.

Any other comments?
No additional comments.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
Each instance in the dataset represents a single Azerbaijani folktale. The folktales are written narrative texts that originate from traditional Azerbaijani oral literature and have been transcribed and published on Azerbaijani Wikisource.

How many instances are there in total (of each type, if appropriate)?
The dataset contains a total of 160 instances. Each instance corresponds to a single Azerbaijani folktale. There is only one type of instance in the dataset.

What was, e.g., the age or gender of the text's authors?
The dataset does not include information about the age, gender, or other demographics of the authors. Many source texts on Azerbaijani Wikisource do not provide author information, and such metadata is largely missing from the dataset.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
The dataset does not contain all possible Azerbaijani folktales. Instead, it represents a non-random sample of folktales collected from Azerbaijani Wikisource. The larger set consists of the full body of Azerbaijani folk narratives, including oral traditions, printed folklore collections, and unpublished or region-specific variants that are not available in digital form. The dataset is not guaranteed to be fully representative of the entire spectrum of Azerbaijani folktales in terms of regional variation, historical period, or narrative style. The selection is constrained by the availability of texts in digital and openly licensed form on Wikisource. No formal validation of representativeness was performed. Many Azerbaijani folktales exist only in oral form or in copyrighted print collections and were therefore unavailable for inclusion. As a result, the dataset reflects only a subset of the larger folklore tradition.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
Each instance consists of raw Azerbaijani narrative text (a folk tale), accompanied by basic metadata and simple

quantitative features such as word and character counts.

Is there a label or target associated with each instance? If so, please provide a description.

There is not a label or target associated with each instance.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Some metadata fields (e.g., author, date, region, and thematic labels) are missing from individual instances due to their absence in the original sources; no content was intentionally removed or redacted.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

"No, the dataset does not include relationships between individual instances; each instance is independent.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

No explicit relationships between individual instances are provided; all instances are treated as independent data points.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

The dataset may include minor formatting noise, spelling variation, and thematically similar or formulaic tales, but no known systematic errors or deliberate duplication.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is mostly self-contained in terms of usable text and metadata. However, it relies on external URLs for provenance. There is no guarantee that those URLs remain valid over time, no official archive of the original content, and potential access or copyright restrictions depending on each source site.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

The dataset does not contain confidential data; all content is derived from publicly accessible narrative texts and does not involve private or protected

communications.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
Some instances may include violent, frightening, or culturally dated themes typical of folk narratives, which could be upsetting to some readers if viewed directly.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
No, the dataset does not relate to real, identifiable people.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
No, the dataset does not contain information about individuals and does not identify any subpopulations.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
"No, it is not possible to identify any individuals, either directly or indirectly, from the dataset.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
"No, the dataset does not contain any data that could be considered sensitive or personally identifiable.

Any other comments?
No additional comments.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
The data is directly observable text scraped from az.wikisource.org. It is not self-reported or inferred. Validation comes from Wikisource's community editing and archival process, not from formal verification.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?
Python script (e.g., with libraries like requests and BeautifulSoup or Scrapy) is used to scrape tale content and metadata from az.wikisource.org. This is a programmatic, automated process that extracted text, titles, source URLs.

How big is the data? If it is a subsample how was it sampled? Was the data collected with consent?

How was the data pre-processed, and what metadata is available?
The dataset contains 160 folktales. The folktales were collected as a non-random sample from Azerbaijani Wikisource and are not representative of all Azerbaijani folktales due to digital availability constraints. Consent is not applicable because the data comes from publicly available, licensed texts. No preprocessing was performed; the dataset contains raw text with basic metadata, including title, source URL. Some metadata fields, such as author, date, and region, are missing.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
Based on the available dataset, there is no explicit indication that it is a sample from a larger collection.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
The data was collected by Nijat Jafarov and Namig Planov as part of a university course project. No external workers or contractors were involved, and no monetary compensation was provided.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.
The data was collected by the authors over the course of the university semester (specify exact months if needed, e.g., September–December 2025). The data instances were created during the same timeframe, as they were generated or gathered specifically for course projects.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
No formal ethical review processes (e.g., by an institutional review board) were conducted for this dataset, as the data collection was performed solely by the authors for a university course project and did not involve sensitive personal information or external participants.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
No, the dataset does not relate to people.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or

show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Any other comments?
No additional comments.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
No preprocessing, cleaning, or labeling was performed on the data; it was collected and stored in its original form.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
Yes, the raw data was saved. No preprocessing, cleaning, or labeling was performed, so the saved data is the same as the original collected data.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
No, the software is not used for this purpose.

Any other comments?
No additional comments.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.
Yes, the dataset has been used for tasks including tokenization and spell checking as part of the course project.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
Yes, there is a repository that links to the project that use this dataset.

What (other) tasks could the dataset be used for?
The dataset could potentially be used for other course-related tasks, including exercises or assignments.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is

there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

Are there tasks for which the dataset should not be used? If so, please provide a description.

The dataset was collected and prepared solely for a university course project, and no preprocessing, cleaning, or labeling was performed. As such, the dataset may contain inconsistencies or errors, and it should not be used for high-stakes applications or decision-making that could impact individuals. Since it does not contain personal or sensitive information, the main concern is technical quality. Future users should review the data carefully and apply appropriate validation or cleaning before using it for any further research or experiments.

Any other comments?

No additional comments.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, the dataset may be distributed for public use. Distribution will be handled by the authors, and no restrictions are placed on access beyond standard academic or public sharing practices.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

he dataset is expected to be distributed publicly via platforms such as HuggingFace and potentially GitHub. No digital object identifier (DOI) has been assigned at this time.

When will the dataset be distributed?

The dataset will be distributed after the completion and assessment of the university course project, subject to the instructor's approval.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset will be distributed by the authors under a standard open academic use license (e.g., CC BY 4.0) for non-commercial research and educational purposes. No fees are associated with accessing the dataset. Exact licensing terms will be provided alongside the dataset upon public release.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No third parties have imposed intellectual property or other restrictions on the data. All data was collected and created by the authors for the course project.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
No export controls or other regulatory restrictions apply to the dataset or its individual instances.

Any other comments?
No additional comments.

Maintenance

Who will be supporting/hosting/maintaining the dataset?
The dataset will be hosted and maintained by the authors (Nijat Jafarov and Namig Planov) for the duration of the university course. Support will be provided on a best-effort basis if needed within the scope of the course project.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
The dataset owners and curators can be contacted via email: njafarov24614@ada.edu.az , nplanov24625@ada.edu.az

Is there an erratum? If so, please provide a link or other access point.
No erratum has been issued for this dataset.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
The dataset is not expected to be updated. Any corrections or additions would be made by the authors on a best-effort basis during the course, but no formal update or communication mechanism is established.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
The dataset does not contain personal or identifiable information about individuals. Therefore, no specific data retention limits or deletion policies apply.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
Support will be provided on a best-effort basis if needed within the scope of the course project.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
There is currently no formal mechanism for external contributors to ex-

tend or augment the dataset. Any modifications or additions would need to be coordinated directly with the authors. Contributions will not be formally validated, as the dataset was created solely for the purposes of the university course, and no public distribution process is established.

Any other comments?

No additional comments.