DATASHEET:
SOCIAL-SUM-MAL

Rahul Raj M
College of Engineering
Trivandrum
d-tve21jul003@cet.ac.in

Dr Dhanya S Pankaj
College of Engineering
Trivandrum
dhanyaspankaj@cet.ac.in

This document is based on *Datasheets for Datasets* by Gebru *et al.* [1]. Please see the most updated version here.

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.
The dataset was created with a view to assist abstractive text summarization tasks in the Malayalam language. The need for an open and reliable dataset in this field led to this work.

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 as a part of the research work done by Mr. Rahul Raj. M under the supervision of Dr. Dhanya. S. Pankaj at the college of engineering Trivandrum, Affliated to APJ Abdul Kalam Technological University, Thiruvananthapuram, Kerala India.

What support was needed to make this dataset? (e.g.who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)
This project was carried out by the research scholar, and supervisor with the help of a language expert. The scholar was availing AICTE Doctoral Fellowship awarded by the AICTE, Government of India.

Any other comments?
The dataset can not only be used for text summarization, but it is capable of evaluating machine learning models for text classification and paragraph heading generation.

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.
The input is a set of Malayalam text paragraphs collected from the Kerala SCERT school-level textbook. For future usage, the chapter number and paragraph number along with individual input IDs were provided. The annotation tasks include heading generation, multi-sentence summary creation, extreme summarization, question and answer generation, and topic identification.

How many instances are there in total (of each type, if appropriate)?
2000 data samples were annotated manually. For the sake of data augmentation, back translation and paraphrasing were performed. In total 6000 samples were there in 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 contains all possible instances of the sample.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
The data was collected from PDF files. Later, PDF-to-text conversion was done with the necessary tools. As a result, we got raw texts that contained spacing issues, spelling errors, and random noises of encryption.

Is there a label or target associated with each instance? If so, please provide a description.
Yes. Each input paragraph is associated with an identifier (Id) along with the chapter number (chapter_no) and paragraph number (para_no). Corresponding to each input, long summary (long_summary), extreme summary (extreme_summary), a query (query), answer to the query(answer_summary) also the topic of the input (topic) were 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.
The data was collected from the Kerala SCERT Social Science textbook. There exists no such missing information.

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.
N/A

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the the rationale behind them.
No.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
The dataset was collected from PDF documents. The encrypted pdf data has been converted to image format followed by image recognition. We have tried to reduce the typos and grammatical errors as much as possible, also the duplicate data was removed through proper coding.

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.
a)Yes. The data will be available in online permanent storage under the ownership of the authors. So that data availability can be assured. b) The later versions of the dataset (created by authors) will be available on the same platform along with the old versions. The data was collected from SCERT textbooks which are freely available online. c) No licenses or fee restrictions will be for the dataset. As they were created for educational purposes, using the same for dataset generation will not result in any copyright infringements.

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.
No. The data source is the textbooks created for educational purposes, so confidential data won't be present. Individual details might have been represented by disguising the names.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
Textbook data are non-offensive in nature. Especially, in this case, social science data provided in textbooks ensures teaching of civic values to students.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
Data partially related to people. The history portions describes the life and view points of influential people including polititians, freedom fighters, and social workers. For achieving better understanding, some of the concepts were described with examples. These examples include human names who have no existence in reality.

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 identify the population.

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.
We can identify political leaders, social workers, Technicians, artists and other influential people throught this 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.
As the data describes the life and viewpoints of public figures some sensitive data about them such as racial or ethnic origins, religious beliefs, political opinions, and locations were revealed through this dataset.

Any other comments?
No

COLLECTION

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 was directly available in the PDF format. As the data source is the Kerala state government-approved textbook, fact-checks, and further validations are not done.

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. Finally, list when the dataset was first published.
Data was collected within a time span of three months. Annotation and data cleaning took a further seven months.

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?
The data collection is done as a combination of software and human efforts. The textbook in PDF format was converted into images using online tools and character recognition was achieved by the facilities provided by Google. Thereafter summary creation was performed manually.

What was the resource cost of collecting the data? (e.g. what were the required computational resources, the associated financial costs, and energy consumption - estimate the carbon footprint. See Strubell *et al.*[2] for approaches in this area.)
The data collection required computational resources such as text processing software, grammar, and spell check tools as well as a Malayalam editor. This work has been carried out as a part of research work funded by AICTE (All India Council of Technical Education). So that all financial funding were contributed by AICTE. There wasn't any financial burden to bear as the contributors to dataset creation worked voluntarily.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
The dataset is self-contained thus no sampling was performed.

Who was involved in the data collection process (e.g., students, crowd workers, contractors) and how were they compensated (e.g., how much were crowd workers paid)?
The research scholar, research supervisor, and Malayalam language expert contributed to the dataset creation.

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.
The language expert critically evaluated the dataset and the response is provided in the journal itself.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.
Human examples were provided in the data, but the dataset was not solely about humans.

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

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.
NA

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.
NA

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)
NA

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.
NA

Any other comments?
NO

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.
A couple of preprocessing tasks like spelling correction, white space removal, missing character identification, and

paragraph and sentence segmentation were carried out. All these were performed manually by the annotators.

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.
The raw data has been stored in both pdf and image format, out of which textual contents were extracted from the image files. The web link to the raw resource is given here. https://drive.google.com/drive/folders/1nanfmmGwaRz3bq_Ht0oR7pjTjS7QuiU-?usp=share_link

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
Pdf splitter: https://www.ilovepdf.com/split_pdf
Google Input Tool: https://www.google.com/intl/ml/inputtools/try/
Malayalam spelling checker: https://www.stars21.com/spelling/malayalam/
Google docs: https://docs.google.com/

Any other comments?
No.

USES

Has the dataset been used for any tasks already? If so, please provide a description.
No. The dataset has been created recently.

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.
No.

What (other) tasks could the dataset be used for?
It can be used for chapter and multi-document summarization. Paragraph classification and topic identification can also be performed.

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?
The paragraphs were collected in an ordered manner from textbooks so that paragraph numbers and chapter numbers were added as features of the dataset. This mechanism helps to identify paragraphs with the same topic and the contents stick together as a chapter. These additional feature helps to train the model for multi-document summarization, clustering, and order prediction.

Are there tasks for which the dataset should not be used? If so, please provide a description.
The dataset cannot be used for machine translation and sentiment analysis.

Any other comments?
No.

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.
This dataset has been created due to the excessive need for benchmark datasets in Malayalam. So we would like to make this open so that fellow researchers in the area can make use of the same for future works.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
The dataset will be available online through the platform Hugging Face for free of cost

When will the dataset be distributed?
The dataset has been created as an objective of research work. It will be made available to everyone soon after completing the publishment of the research paper.

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.
No. The dataset will be available in free and open to modifying form. However, those who are using the dataset should cite the journal paper in their work.

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.

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.

Any other comments?
No.

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?
The dataset will be available on the Hugging face platform. So they are the people who are going to maintain the data.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
The contact details of the authors such as institutional address and email were provided along with the dataset in the platform so that all queries related to the dataset creation can be addressed. The same information is also available in the journal whose link will also be available in the platform.

Is there an erratum? If so, please provide a link or other access point.
No.

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)?
We are expecting the updation of the dataset in two manners, one by ourselves and the other by user contribution. The frequency of both is unpredictable but the dataset revision by authors will be available on the platform with the release date and other metadata. The modified dataset in collaboration with the third-party curator is strictly based on their willingness to publish the same in the author's platform.

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.
N/A

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.
Yes. Older versions of the data will be archived in the same platform.

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.
Community collaboration is much encouraged in dataset updation. Those who have created new versions of the dataset can communicate the same through the instructions given in the platform. After scrutiny, the dataset will be updated in the platform.

Any other comments?
No.

REFERENCES

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for Datasets. *arXiv:1803.09010 [cs]*, January 2020.

[2] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP. *arXiv:1906.02243 [cs]*, June 2019.