Datasheet for dataset "TIE"

Questions from the Datasheets for Datasets paper, v7.

Jump to section:

Motivation

The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests.

For what purpose was the dataset created?

The dataset was created to address the scarcity of technical domain data within the Indian context for the purpose of auditing and mitigating bias in State-of-the-Art Automatic Speech Recognition (ASR) systems.

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

The videos of dataset has been sourced from NPTEL channel of YouTube and annotated and curated by team of researchers (Anand Kumar Rai, Siddharth D Jaiswal under supervision of Prof. Animesh Mukherjee at IIT Kharagpur.

Who funded the creation of the dataset?

The creation of the dataset was not financially supported by any external funding.

Any other comments?

Composition

Most of these questions are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for specific tasks. The answers to some of these questions reveal information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

The instances within the dataset represent technical lecture videos, encompassing 8740 hours of speech delivered by 332 Indian speakers across more than 20 varied lecture topics. Each video file has been enriched with annotations detailing demographic attributes such as teaching experience, gender, caste, and native region of the respective speaker, as well as audio metadata including speech rate, discipline, and lecture topic.The dataset also contains transcripts generated from YouTube and Whisper ASR for those lecture videos.

How many instances are there in total (of each type, if appropriate)?

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

The dataset represents a sample of instances rather than encompassing all possible instances. The sample is not necessarily random but is rather curated to reflect certain demographics and characteristics, particularly those observed in higher educational institutes in India. Notably, there is a significant skew towards male speakers (approximately 95%) and those belonging to the unreserved caste category (approximately 73%), which mirrors the distribution observed in Indian higher education institutions. Additionally, there are insights into the representation of females in teaching positions and the distribution of lecturers belonging to reserved caste categories. The dataset also reflects the distribution of speakers based on teaching experience and native regions, with certain regions being underrepresented. Furthermore, annotations such as speech rate and discipline are provided for each lecture, which may result in an overlap of speakers across subcategories within the dataset.

What data does each instance consist of?

Each instance within the dataset consists of a technical lecture video delivered by an Indian speaker. Specifically, the data for each instance includes:

- Speech Content: The main content of the lecture delivered by the speaker.
- Speaker Information: Demographic attributes such as gender, teaching experience, caste, native region, and any other relevant information about the speaker.
- Audio Metadata: Information pertaining to the audio characteristics of the lecture, such as speech rate.
- Lecture Metadata: Details regarding the discipline and topic of the lecture.
- ASR Generated Transcripts: Transcripts corresponding to each lecture transcribed by YouTube ASR and OpenAI Whisper

Is there a label or target associated with each instance?

The provided information doesn't explicitly mention a label or target associated with each instance. However, depending on the specific use case or analysis, it's possible that labels or targets could be derived or assigned based on the content or characteristics of the lectures. For example, if the dataset is used for speech recognition tasks, the target could be transcriptions of the speech content. If the dataset is used for demographic analysis, the target could be the demographic attributes of the speakers. If further annotation or labeling is required, it would typically be mentioned in the dataset documentation or specifications.

Is any information missing from individual instances?

The TIE dataset includes comprehensive information for each individual instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?

Each instance represent an individual technical lecture delivered by a speaker, with associated metadata and annotations and no explicit relationships have been mentioned in the dataset.

Are there recommended data splits (e.g., training, development/validation, testing)?

No

Are there any errors, sources of noise, or redundancies in the dataset?

Firstly, the accuracy of manual transcripts may be compromised by human error during transcription. Additionally, self-annotated labels, such as gender and caste, pose challenges as they are not publicly available, making their determination difficult. Moreover, automatic identification of these attributes from names may introduce potential inaccuracies. Furthermore, the formal speech style and use of technical jargon in the videos may impact the performance of Automatic Speech Recognition (ASR) systems. Lectures involving live mathematical derivations or problem-solving, as opposed to those delivered using slides, may also influence ASR performance. Additionally, the non-deterministic outputs of ASR model components may lead to varying transcripts for the same video, further complicating analysis.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset relies on external resources as the videos comprising the dataset have been sourced from the NPTEL channel on YouTube. This implies that the dataset is not entirely self-contained but is instead dependent on the availability and accessibility of the NPTEL channel's content on YouTube. Consequently, any changes, updates, or removals of videos from the NPTEL channel may directly impact the dataset's content and availability. Therefore, users should consider this dependency when utilizing the dataset for analysis or research purposes.

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)?

No, the videos have been sourced from YouTube which are publicly availabe and the speaker identity information has been anonymized in the annotated dataset.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No

Does the dataset relate to people?

Yes, the dataset relates to people as it consists of technical lecture videos delivered by 332 Indian speakers. These speakers are individuals who deliver lectures on various topics covered in the dataset. Additionally, demographic attributes such as gender, caste, teaching experience, and native region are annotated for each speaker, further establishing the relationship between the dataset and individuals. Therefore, the dataset involves people in the context of delivering lectures and associated demographic information.

Does the dataset identify any subpopulations (e.g., by age, gender)?

Yes, the dataset identifies subpopulations based on various demographic attributes. Specifically, it annotates demographic information such as gender, teaching experience, caste, and native region for each speaker. This allows for the identification of subpopulations within the dataset based on these attributes

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?

The dataset does not allow for the direct identification of individuals. The dataset consists of technical lecture videos delivered by multiple speakers, with demographic attributes such as gender, teaching experience, caste, and native region annotated for each speaker. While this information provides insight into the characteristics of the speakers, it may not be sufficient on its own to directly identify individuals. However, indirect identification of individuals could potentially occur if the dataset is combined with other external data sources like personal websites of lecturers.

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)?

No

Any other comments?

Collection process

[T]he answers to questions here may provide information that allow others to reconstruct the dataset without access to it.

How was the data associated with each instance acquired?

The data associated with each instance, which comprises technical lecture videos delivered by Indian speakers along with demographic attributes and metadata, have been acquired from the NPTEL (National Programme on Technology Enhanced Learning) channel on YouTube. Additionally, demographic attributes such as gender, teaching experience, caste, and native region may have been collected through manual annotation or self-reporting by the speakers themselves. The acquisition process involved a combination of data collection from publicly available sources and manual annotation to enrich the dataset with relevant metadata.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

The TIE dataset was compiled through a systematic process combining manual human curation and software programs. Initially, metadata for courses, encompassing details like course name, discipline, institute, instructor, and course weblink, was scrapped from the NPTEL website using python script, resulting in a corpus of 2567 courses. Subsequently, titles and weblinks for associated lecture videos, averaging approximately 31 per course, were also gathered, resulting in a comprehensive dataset comprising 78,222 lecture videos. In the following phase, a script was utilized to filter out videos hosted on YouTube and possessing ground truth transcripts, yielding a refined dataset of 9860 lecture videos alongside all previously acquired metadata. The dataset preparation process concluded with the acquisition of video files, followed by the extraction of their audio tracks into MP3 format using software tools. Additionally, demographic attributes were annotated manually, contributing to the finalization of the dataset.

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 sample of larger set of videos as mentioned in previous answer. The dataset compilation process follows a deterministic sampling strategy rather than a probabilistic one. Videos are selected based on specific criteria i.e their availability on YouTube and the presence of corresponding ground truth transcripts.

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 videos of dataset has been sourced from NPTEL channel of YouTube and annotated and curated by team of researchers (Anand Kumar Rai, Siddharth D Jaiswal under supervision of Prof. Animesh Mukherjee at IIT Kharagpur.

Over what timeframe was the data collected?

The creation timeframe for the data associated with the instances in the dataset involved approximately 5 months. This timeframe encompassed several tasks, including scraping video metadata from the NPTEL website, transcribing the videos using YouTube and Whisper ASR, and annotating attributes corresponding to speakers.

Were any ethical review processes conducted (e.g., by an institutional review board)?

No

Does the dataset relate to people?

Yes, the dataset relates to people as it consists of technical lecture videos delivered by Indian speakers. These speakers are individuals who deliver lectures on various topics covered in the dataset. Additionally, demographic attributes such as gender, teaching experience, caste, and native region are annotated for each speaker, further establishing the relationship between the dataset and individuals. Therefore, the dataset involves people in the context of delivering lectures and associated demographic information.

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

The demographic attributes and other metadata were manually annotated, with assistance from information available on academic institutions' websites or other publicly accessible sources.

Were the individuals in question notified about the data collection?

No, since the data was sourced from individual lectures publicly available on YouTube and their metadata on publicly accessible sources.

Did the individuals in question consent to the collection and use of their data?

NA

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

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?

NA

Any other comments?

Preprocessing/cleaning/labeling

The questions in this section are intended to provide dataset consumers with the information they need to determine whether the "raw" data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a "bag-of-words" is not suitable for tasks involving word order.

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)?

Yes, preprocessing, cleaning, and labeling of the data were conducted as part of the dataset creation process. The prepocessing and labelling steps includes the following:

- Scraping video metadata from the NPTEL website: This likely involved extracting relevant information such as course names, disciplines, instructors, and weblinks.
- Transcribing the videos using YouTube and Whisper ASR: This process converts the spoken content of the lecture videos into textual transcripts.
- Annotating attributes corresponding to speakers: Demographic attributes such as gender, teaching experience, caste, and native region were manually annotated for each speaker.

These steps involve various forms of preprocessing and cleaning, such as extracting metadata, transcribing audio content, and annotating demographic attributes. Additionally, any missing values or inconsistencies in the data may have been addressed during these preprocessing stages to ensure the quality and completeness of the dataset.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

Yes, the ground-truth and ASR generated transcripts on which audit was conducted is available for any other unanticipated usage.

Is the software used to preprocess/clean/label the instances available?

The script used for downloading video from YouTube extracting mp3 and ground-truth transcript is available

Any other comments?

Uses

These questions are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

Has the dataset been used for any tasks already?

Not yet.

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.

What (other) tasks could the dataset be used for?

The dataset presents a versatile resource with potential applications across various domains beyond its primary focus on automatic speech recognition and demographic analysis. With its extensive collection of technical lecture videos and associated metadata, researchers can explore tasks such as speaker identification, natural language processing, education research, bias mitigation in AI systems, and language learning. Additionally, the annotated demographic attributes offer opportunities for demographic analysis and understanding representation in technical education. The dataset's content and richness make it valuable for developing and evaluating systems in speech processing, education technology, and beyond. Its availability opens avenues for diverse research endeavors aimed at advancing understanding and innovation in related fields.

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

The composition, collection, and preprocessing of the dataset may have implications for its future uses. Biases in demographic representation, such as the overrepresentation of certain groups or regions and underrepresentation of others, could impact the generalizability of findings and the fairness of applications built on the data. Additionally, the accuracy and consistency of transcriptions, as well as the completeness and integrity of metadata, may influence the effectiveness of downstream analyses or models trained on the dataset. Ethical considerations, including privacy protection for individuals represented in the data, are also paramount. Researchers and users should be mindful of these factors and conduct thorough validation, preprocessing, and documentation to ensure the reliability and ethical use of the dataset in future endeavo

Are there tasks for which the dataset should not be used?

If so, please provide a description.

Any other 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?

Yes, the dataset has been made publicly available

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?

The dataset link is available

When will the dataset be distributed?

The dataset is already available

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

The dataset is distributed under license Attribution-ShareAlike 2.0 Generic (CC BY-SA 2.0).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

No

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

No

Any other comments?

Maintenance

These questions are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan with dataset consumers.

Who is supporting/hosting/maintaining the dataset?

Anand Kumar Rai is one of the researchers involved in the creation of the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

For any inquiries , feedback, or collaboration on the dataset, please contact [raianand.1991@gmail.com].

Is there an erratum?

No

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

If necessary, updates will be communicated through the dataset's GitHub repository and via email to the mailing list of dataset users.

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)?

No

Will older versions of the dataset continue to be supported/hosted/maintained?

Yes

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

For any inquiries on collaboration on the dataset, please contact [raianand.1991@gmail.com].

Any other comments?