DASP Project

Group 2A - Argumentation in Author-Reviewer Paper Discussions

License for the Data

There exists no license for the data by OpenReview as far as we are aware.

Neither https://openreview.net/terms nor https://openreview.net/about give any licensing information.

Data Statement

Terms:

Speaker: individual who produced linguistic behavior included in the dataset
Annotator: people who annotate the raw data
Curator: people who are involved in the selection of which data to include
Stakeholders: Direct, Indirect
Bias: systematic and unfair discrimination, pre-existing bias (society), technical bias (algorithm decisions), emergent bias (applied to different context)

According to https://www.aclweb.org/anthology/Q18-1041.pdf a full data statement should consist of the following paragraphs:

A. **CURATION RATIONALE**
Which texts were included and what were the goals in selecting texts

B. **LANGUAGE VARIETY**
Languages differ from each other in structural ways that can interact with NLP algorithms, include language tag (BCP-47) and prose description

C. **SPEAKER DEMOGRAPHIC**
Native or non-native speaker (L1, L2), age, gender, race/ ethnicity, native language, socioeconomic status, number of different speakers represented, presence of disordered speech

D. **ANNOTATOR DEMOGRAPHIC**
Includes age, gender, race/ ethnicity, native language, socioeconomic status, training in linguistics/other relevant discipline (covers guideline developers as well)

E. **SPEECH SITUATION**
To better understand the context, includes time and place, modality (spoken, written), scripted/ edited vs. spontaneous, synchronous/ asynchronous interaction, intended audience

F. **TEXT CHARACTERISTICS**
Genre and topic of the texts

G. **RECORDING QUALITY**
For audiovisual recordings

H. **OTHER**
Other information of relevance

I. **PROVENANCE APPENDIX**
For datasets built out of existing datasets

Curation Rationale

The dataset was created in order to analyze how revisions of submitted scientific papers for a conference (hosted on OpenReview.net) change over time. Data was crawled as part of a student project at TU Darmstadt. The scope and level of detail was defined by the project owner. The goal was to map comments and reviews posted on a submission to its revisions.

Language Variety

OpenReview hosts scientific papers for conferences and workshops. The language in those papers is English (en-US, en-GB or ESL). The comments on the submissions are also posted in English (en-US, en-GB or ESL), but may contain colloquialism, bullet points and fragments of sentences.

Speaker Demographic

Speakers are scientists attempting to publish in various venues. We can assume an on average higher socioeconomic status from this.
Speakers can be from any country. Specific information is not available.
Age information is not available except that speakers are at least adults.
Papers can be written by multiple people. Comments and reviews are usually written by only one person.

Annotator Demographic

The data was annotated by an algorithm. The algorithm was written by a male student at the TU Darmstadt.

Formatiert: Zeilenabstand: Mehrere 1.08 ze

Speech Situation

All data was retrieved in written form on selected venues between 2013 and 2020.
Submissions are scientific papers.
Comments and reviews are written text ranging in length from one sentence to multiple paragraphs.
Reviews are anonymous and address the content of the submission. They can include informal language constructs, questions, suggestions and advice. Venues have different standards in handling reviews, but they predominantly include a conclusion in form of a rating and the confidence of the reviewer in the review's correctness. A submission usually has a fixed number of reviews by specifically assigned reviewers. The number of reviews is decided by the venues; three is a common choice.
Comments are often a direct (possibly anonymous) reply to a review and may contain informal language. There can also be comments on the submission that are no reviews, but simple questions or remarks. Comments can be written by the paper authors, reviewers or other people.

Text Characteristics

Submissions were posted with the intention of being accepted to a venue. Reviews are intended to provide feedback for the author, to suggest revisions and to aid the decision whether a submission should be rejected or accepted.

Recording Quality

N/A

Other	-
Provenance Appendix	N/A

Formatiert: Zeilenabstand: Mehrere 1.08 ze