Datasheet

1 Datasheet

1.1 Motivation

1. For what purpose was the dataset created?
   Our dataset was created to address the novel Emotion Forecasting task, which aims to transform affective forecasting into a deep learning problem by modeling it through data on two-party interactions.

2. Who created the dataset and on behalf of which entity?
   It will be released after the paper review.

3. Who funded the creation of the dataset?
   It will be released after the paper review.

1.2 Distribution

1. 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 is open to the public.

2. How will the dataset be distributed (e.g., tarball on website, API, GitHub)?
   The dataset will be distributed through Google Drive and the code used for developing baseline models through GitHub.

3. Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
   No.

4. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?
   No.

1.3 Maintenance

1. Who will be supporting/hosting/maintaining the dataset?
   The authors of the paper will support, host, and maintain the dataset.

2. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
   The owner/curator/manager(s) of the dataset can be contacted through emails and will be released after the paper review.

3. Is there an erratum?
   No. If errors are found in the future, we will release errata on the main web page for the dataset (`https://github.com/Anonymize-Author/Hi-EF`).

4. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?
   Yes, the datasets will be updated whenever necessary to ensure accuracy, and announcements will be made accordingly. These updates will be posted on the main web page for the dataset (`https://github.com/Anonymize-Author/Hi-EF`).

5. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted?)
   N/A

6. Will older versions of the dataset continue to be supported/hosted/maintained?
   Yes, older versions of the dataset will continue to be maintained and hosted.

7. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?
   No.

1.4 Composition

1. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries?)
   Each video clip represents a snapshot of a speaker's interaction, including video ID, utterance text, facial movements, scene, interaction polarity, intensity, emotion, and emotion uncertainty.

2. How many instances are there in total (of each type, if appropriate)?
   The dataset (Hi-EF) includes 3,069 MCIS and 5,242 annotated video clips.

3. Does the dataset contain all possible instances or is it a sample of instances from a larger set?
   The datasets contain possible instances.

4. Is there a label or target associated with each instance?
   Yes, each instance includes both input and target (prediction) variables.

5. Is any information missing from individual instances?
   No.

6. Are there recommended data splits (e.g., training, development/validation, testing)?
   We partitioned the dataset, comprising 3,069 MCIS, into training (70%) and testing (30%) sets, with the training set further segmented into a validation subset.

7. Are there any errors, sources of noise, or redundancies in the dataset?
   There is noise present in the dataset due to the subjectivity involved in emotion annotation, which may lead to some bias in the labels. To address this issue, we introduced an "emotion uncertainty" variable to describe and mitigate the impact of such subjectivity on the noise.

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

9. Does the dataset contain data that might be considered confidential?
   No.

10. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
    No.

1.5 Collection Process

1. How was the data associated with each instance acquired?
   The data associated with each instance was acquired by clipping long videos from a series of TV dramas according to timestamps.

2. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?
   We used CPUs and script programs to trim video files based on subtitle files.

3. 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)?
   We hired JD Crowdsourcing to perform the annotation for us, and the payment for annotating video samples was calculated based on each individual's working hours and hourly wage.

4. Does the dataset relate to people?
   No.

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

1.6 Uses

1. Has the dataset been used for any tasks already?
   No, this dataset has not been used for any tasks yet.

2. What (other) tasks could the dataset be used for?
   Our dataset is designed for the Emotion Forecasting task; for derived tasks, please refer to Section E in the Appendix.

3. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?
   Our current dataset is built based on dialogues from TV dramas. In the next version, we plan to collect data from real-world scenarios to expand the dataset. Any changes in the next version and updates to the user guidelines will be documented and shared through the dataset webpage (`https://github.com/Anonymize-Author/Hi-EF`).

4. Are there tasks for which the dataset should not be used?
   No.