Datasheet: *DAIC-WOZ Depression Database*

Author: *Ruotong Gao*

Organization: *School of Information, University of Michigan*

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.

1. **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 data (interviews) were collected as part of a larger effort to create a computer agent that interviews people and identifies verbal and nonverbal indicators of mental illness. The corpus has been used to support the creation of an automated interviewer agent, and for research on the automatic identification of psychological distress.

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

   The creators of this dataset are: Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency, on behalf of USC Institute for Creative Technologies.

3. **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 work is supported by DARPA under contract W911NF-04-D-0005 and by the U.S. Army RDECOM.

4. **Any other comments?**

   None.

Composition

Dataset creators should read through the questions in this section prior to any data collection and then provide answers once collection is complete. 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.

1. **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.

   This database is part of a larger corpus, the Distress Analysis Interview Corpus (DAIC), that contains clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder. Data collected include audio and video recordings and extensive questionnaire responses. Further, they provide the scores that every individual provided on the PHQ-8 depression inventory.

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

   There are 189 instances (participants) in total.

3. **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).

   This dataset contains a sample of instances: participants are drawn from a population living in the Greater Los Angeles metropolitan area – from the general public. No tests were run to determine representativeness.

4. **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: audio recordings - raw audio and transcripts of the interview are being provided, allowing the participants to compute additional features on their own; audio features - audio features are extracted using the COVAREP toolbox (v. 1.3.2) available at: https://github.com/covarep/covarep; questionnaire data - participants completed PHQ-8 questionnaires prior to the interview, including basic demographic questions, established measures of psychological distress, and a measure of current mood; video features - based on the

OpenFace framework, it provides different types of video features include facial landmarks, facial landmarks, gaze, head pose, and action units (AUs).

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

   Yes. The level of depression is labelled with a PHQ-8 binary label (PHQ8 Scores >= 10), PHQ8 Scores, and single responses for every question of the PHQ8 questionnaire per recording.

6. **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.

   Yes. Session 373, 444, 451, 458, 480 contain technical issues which may influence analysis effectiveness.

7. **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.

8. **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.

   Yes. There are recommended data splits include training, developing, and testing sets. However, the authors didn't indicate the rationale behind such splits.

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

   Not indicated.

10. **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.

    This dataset is self-contained.

11. **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.

Yes. The dataset includes audio recordings, self-assessed subjective depression questionnaire results, and facial features of participants, which may be considered confidential and protected by legal privilege and doctor-patient confidentiality.

12. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.

    Yes. Due to the depression diagnosis nature of this dataset, it may trigger depression, anxiety, and potential mental health issues if viewed directly.

13. **Does the dataset relate to people?** If not, you may skip the remaining questions in this section.

    Yes.

14. **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.

    Yes. The dataset identifies participants' binary gender in the training and developing set. A rough analysis of their respective distributions can be found here: https://paperswithcode.com/paper/raw-audio-for-depression-detection-can-be/review/.

15. **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.

    Yes, it is possible to identify individuals indirectly. As the dataset consists of "raw" audio recordings, participants who are from a limited and clearly indicated population may be identified through audio recognition technology.

16. **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.

17. **Any other comments?**

    None.

Collection

As with the previous section, dataset creators should read through these questions prior to any data collection to flag potential issues and then provide answers once collection is complete. In addition to the goals of the prior section, the answers to questions here may provide information that allow others to reconstruct the dataset without access to it.

1. **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 associated with each instance was acquired by Wizard-of-Oz interviews conducted by an animated virtual interviewer called Ellie, controlled by a human interviewer in another room. The PHQ-8 results were acquired by asking participants to complete a subjective self-assessment questionnaire.

2. **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?

   In the Wizard-of-Oz interviews, the participant was recorded by a camera, high-quality close-talking microphone, and Kinect, while the agent was recorded through screen-capture software.

3. **If the dataset is a sample from a larger set, what was the sampling strategy (e.g. deterministic, probabilistic with specific sampling probabilities)?**

   Not indicated.

4. **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 participants were recruited from Craigslist and data annotators were involved in the annotation process. The compensation details are not indicated.

5. **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.

   Not indicated. The dataset was first published in 2014.

6. **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.

*Probably. The authors mentioned ethical review guidelines of their institution, but didn't indicate directly whether ethical review process was conducted. *

7. **Does the dataset relate to people?** If not, you may skip the remainder of the questions in this section.

   Yes.

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

   Yes, the research group collected the data directly and their interviews were all conducted at ICT.

9. **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.

   Yes. Detailed descriptions are not indicated.

10. **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.

    Yes. Participants completed a consent form (which included optional consent that allowed their data to be shared for research purposes).

11. **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).

    Not indicated.

12. **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.

    Not indicated.

13. **Any other comments?**

    The experiment process ended with a "cool-down" phase, to ensure that participants would not leave the interview in a distressed state of mind.

Preprocessing / Cleaning / Labeling

Dataset creators should read through these questions prior to any pre-processing, cleaning, or labeling and then provide answers once these tasks are complete. 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.

1. **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.

   Yes. Data has been transcribed and annotated for a variety of verbal and non-verbal features. All the transcribed interviews were annotated to remove identifying information. Parts of the transcribed corpus have
   been annotated with dialogue-level information to support the development and training of natural language understanding for the agent.

2. **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 are saved in the same folder with the preprocessed/cleaned/labeled data.

3. **Is the software used to preprocess/clean/label the instances available?** If so, please provide a link or other access point.

   Yes. For the audio features the research group utilised COVAREP (v1.3.2), a freely available open source Matlab and Octave toolbox for speech analysis. The video features were collected based on the OpenFace framework.

4. **Any other comments?**

   None.

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.

1. **Has the dataset been used for any tasks already?** If so, please provide a description.

   Yes. The corpus has been used to support the automated agent's interactive capabilities by developing custom acoustic and language models for speech recognition, training classifiers for natural language understanding, and informing the creation of dialogue policies. The corpus has also been used to support the agent's capabilities for distress detection, using multiple types of information including visual signals, voice quality, and dialogue-level features. In addition, it has been used for AVEC 2017 Real-life Depression, and Affect Recognition Workshop and Challenge, aiming at comparison of multimedia processing and machine learning methods for automatic audiovisual depression and emotion analysis.

2. **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.

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

   The corpus has been used to support the creation of an automated interviewer agent, and for research on the automatic identification of psychological distress. This robust dataset has the potential to help various researchers address questions across areas of mental health, human-agent interactions, and verbal and non-verbal behavior.

4. **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 use of data for illustrative purposes in publications is allowed. Publications include both scientific papers and presentations for scientific/educational purposes. In this case, the identity of the subjects should be protected (no release of identifiable information for subjects).

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

   Given that this dataset contains facial features and audio recordings, it shouldn't be used for audio or facial identification which might poses risk on participants' privacy.

6. **Any other comments?**

   None.

Distribution

Dataset creators should provide answers to these questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.

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?** If so, please provide a description.

   Yes, the dataset can be distributed by asking the institution for access. The database is made available for research purposes only. Any commercial use of this data is forbidden.

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

   To be able to download the DAIC-WOZ database, please download, sign and return the agreement form to this e-mail address. Please note that, unfortunately, due to consent constraints we are only allowed to distribute the data to academics and other non-profit researchers. Please use your academic e-mail address when requesting the data download. The DOI is not indicated.

3. **When will the dataset be distributed?**

   The dataset was first released in 2014.

4. **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 crawled data copyright belongs to the authors of the reviews unless otherwise stated. There is no license, and all publications reporting on research using this database have to acknowledge this by citing the following article: Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, Louis-Philippe Morency, "The Distress Analysis Interview Corpus of human and computer interviews", in Proceedings of Language Resources and Evaluation Conference (LREC), 2014.

5. **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.

6. **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.

   Not indicated.

7. **Any other comments?**

   None.

Maintenance

As with the previous section, dataset creators should provide answers to these questions prior to distributing the dataset. These questions are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan with dataset consumers.

1. **Who is supporting/hosting/maintaining the dataset?**

   Institute for Creative Technologies, University of Southern California.

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

   Jill Boberg: boberg@ict.usc.edu.

3. **Is there an erratum?** If so, please provide a link or other access point.

   Not indicated.

4. **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)?

   There is an extended DAIC database which is the extended version of DAIC-WOZ database for depression and PTSD assessment, developed by ICT. The level of depression severity (PHQ-8 questionnaire) in the new database was assessed from audio-visual recordings of US Army veterans interacting with a virtual agent conducting a clinical interview and driven by a human as a Wizard-of-Oz (DAIC-WOZ corpus). The DAIC corpus contains new recordings with the virtual agent being, this time, fully driven by artificial intelligence, i.e., without any human intervention.

5. **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.

No, but the users should sign an agreement form and the ICT Institute at University of Southern California is allowed to change the terms of use at any time. In this case, users will be informed of the changes and will have to sign a new agreement form to keep using the database.

6. **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.

   Not indicated. Currently, two versions of DAIC-WOZ database (2014 & 2019) are all being maintained by ICT.

7. **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.

   Not indicated.

8. **Any other comments?**

   None.

All information in this datasheet is taken from the following sources: https://dcapswoz.ict.usc.edu/; https://dcapswoz.ict.usc.edu/wwwutil_files/DAICWOZDepression_Documentation.pdf; https://aclanthology.org/L14-1421/; https://dl.acm.org/doi/10.1145/3133944.3133953; https://sites.google.com/view/avec2019/home.