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.*

Ethical review processes, apart from the creation of these datasheets, have not been conducted.

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

The dataset has been collected through publicly accessible websites.

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 other- wise reproduce, the exact language of the notification itself.*

Individuals were not notified about the collection. Textual data was scraped without location, author or timestamp information. The author of a piece of text is untracable

Preprocessing/cleaning/labelling

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)?

The following steps were taken to create the dataset:

1. Collecting raw textual data
   a. As discussed in the collection process, two internet platforms were scraped.
2. Gathering a team of annotators.
   a. I found people in my network willing to help me annotate the textual data.
   b. The annotation team consisted of 6 females (ages between 21 and 54), and 7 males (ages between 21 and 60).
3. Set-up of the digital annotation environment.
   a. We used the annotation tool doccano, and served it to the anotators using Google Cloud Platform.
   b. The annotators were presented raw textual data (i.e., no preprocessing was done)
4. One set was annotated by all annotators to determine inter-annotator agreement.
5. All annotators labeled an additional set containing pieces of text.
6. Inter-annotator agreement was measured, and some annotators' annotations were taken out to optimize this metric.

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

Yes, the data is stored, but not yet publicly available.

Uses

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

The dataset has not yet been used.

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

To the best of my knowledge there is no risk for personal harm. In practical terms: One should consider class imbalance before using the data to train a model.

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.*

The dataset will be shared with the company I did my internship with (Tracey). Apart from this, the plan is to make the dataset publicly available to stimulate Dutch emotion recognition research.

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

The plan is to share the dataset either via Github, or an university repository. As this is not certain, there is no DOI associated with it (yet).

When will the dataset be distributed?

I am aiming to make the dataset available by September 2021.

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.*

This is yet to be determined by the entity I do my research with (Jheronimus Academy of Data Science).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? *If so, please describe these restric- tions, 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.*

There are no fees or restrictions.

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.*

To the best of our knowledge: no.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

Unknown

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

All questions and comments can be sent to: jaspernieuwdorp@icloud.com

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

This depends on where the dataset will be hosted / made accessible. In Github, a readme and erratum can be easily included in the repository.

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)?*

All changes to the dataset will be announced through the repository. Potential errors sent to the e-mail address above will be changed in the dataset, if correct.

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.*

There are no restrictions or limmits on the retention of the data.

Will older versions of the dataset continue to be sup- ported/hosted/maintained? *If so, please describe how. If not, please describe how its obsolescence will be communicated to users.*

They will continue to be supported with all information.

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.*

If others want to extend or build on the dataset, there are several things to consider:

1. One should be aware of the emotion framework used to create this dataset. The 8 basic emotions of Robert Plutchik have initially been used to annotate the text in this dataset.
   a. If one is to change the emotion framework, the translation to other emotion categories should be scientifically grounded. Too many datasets with arbitrarily chosen emotion models exist.
2. One should question if a cross-domain setting is the best way to go (i.e., should one add data from different sources to this data?). Perhaps the additional text should come from the same sources as I used (see the data collection process).
3. If one wants to discuss potential additions or augmentations to the dataset, please contact me using the e-mail address above.