Motivation

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 dataset has been created to train emotion-recognition algorithms in detecting emotion from Dutch text.

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

This dataset has been created by Jasper Nieuwdorp, as part of his thesis that served as partial fulfillment of the master in Data Science & Entrepreneurship.

Who funded the creation of the dataset? *If there is an associated grant. please provide the name of the grantor and the grant name and number.*

No funding was granted.

Composition

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 decription.*

The instances in the dataset are pieces of natural text. They represent natural language, written by real people. The pieces of text are labeled with the emotion that is conveyed in the text.

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

The dataset contains 1500 pieces of text. Instances can be full sentences, a group of sentences, or groups of words.

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).*

The dataset contains text, scraped from the platforms Trustpilot & Reddit. As Dutch text was required, it only contains text from Dutch sub-reddits. For trustpilot, pages of a selected number of last-mile transportation companies were scraped. Therefore, this data is a sample of the total amount of textual data that these platforms contain. Representativeness is not validated as the acquisition of Dutch, human-written textual data on itself is a challenging task. Additionally, one could question if emotion is expressed similarly in different domains (Klinger, 2018). Therefore, we focussed on one domain: last-mile parcel transportation, using the Trustpilot reviews. We later added Reddit as we found that reviews were too extremely polarized.

What data does each instance consist of? *"Raw" data (e.g., unpro- cessed text or images)or features? In either case, please provide a de- scription.*

Raw, unprocessed text. Each instance contains text in string format. And the emotion label in string format.

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

Each piece of text is associated with the emotion that is conveyed in the text.

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.*

One should look out for class imbalance, as some emotions might occur more frequently in the dataset than others. Over / undersampling might be required to achieve accurate machine learning performance evaluation.

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

The instances are labeled using a team of annotators. There might be labelling errors present in the dataset.

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.

Klinger, R. (2018, August). An analysis of annotated corpora for emotion classification in text. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 2104-2119).

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient con- fidentiality, data that includes the content of individuals non-public communications)? *If so, please provide a description.*

All data was derived from publicly available sources, using a free-to-the-public API (for Reddit).

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

The instances in the dataset might contain offensive language. The data hasn't been filtered on offensive language, and the text has been scraped off forums with real people that discuss with each other.

Does the dataset relate to people?

Yes, the dataset contains natural language, written by real people.

Does the dataset identify any subpopulations (e.g., by age, gender)? *IIf so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

The dataset doet not identify subpopulations. From the platforms, only the text has been scraped. Not the author, or any information about the author.

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.*

One could google a piece of text from the dataset, and potentially find the username of the author. There exists a potential risk in identification here.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orien- tations, 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)?

This data does not contain any sensitive data.

Collection Process

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 raw review texts were scraped from a company page on Trustpilot. All reviews for the respective companies were scraped. To scrape reddit text, their API was used. The API allows for collecting all posts and comments (and comments on comments etc), for a particular subreddit.

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

The Reddit API that is open to the public, and the python package PRAW was used to make the script. Trustpilot reviews were scraped using Beautifulsoup.

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

From Reddit the following Dutch forums were scraped: Circeltrek, Geschiedenis, papgrappen, RMTK, D66, Politiek, coronanetherlans, eindhoven, roermond, tokkiefeesboek, nederlands, VeganNL. From Trustpilot, reviews from the following companies were scraped: PostNL, DHL, GLS, Ziggo, Trunkrs, ParcelParcel. The strategy here was that we required only Dutch texts, and we're developing the dataset for a company that uses it to track parcels in the e-commerce supply chain.

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)?

Nobody helped in the data collection process. This was only done by Jasper Nieuwdorp (graduate student). There were a team of annotators involved in labeling (annotating) the dataset, but this will be further elaborated in the respective section (Preprocessing/ cleaning/ labelling).

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 time- frame in which the data associated with the instances was created.*

Unknown, complete subreddits / company Trustpilot pages were scraped.