(Legacy) Final Dataset Report

This template is created according to Gebru, Timnit, et al. "Datasheets for datasets." *Communications of the ACM* 64.12 (2021): 86-92. You can visit the paper here or read it on ODTUClass: Datasheets for Datasets

Date Updated: 30 Oct 2022

Last Updated by: @ Mert Mecit

Team Name: WhiteRabbit

Team Members: @ Furkan Bol @ Selin Deniz @ Gizem Kaya @ Mert Mecit

Dataset Name: Final Dataset

Motivation

The following questions are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests. The latter may be particularly relevant for datasets created for research purposes.

	Question	Answer
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?	This Final Data is created to combine price and news data into one single data that enables us to train our ensemble models.
2	**Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?**	The dataset is collected for Project Rabbidity by the WhiteRabbit team under Wonderland Co. The project aims to provide a decision support system for investors of Brent Crude.
3	**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.	Every step of the project, including the datasets, are funded by Wonderland Co.
4	**Any other comments?**	There is a “final_data” csv file used for analysis and modelling now. This dataset is uploaded to Bitbucket so the data version control is assured that way. In Sprint 3, it is planned to use DVC.

Composition

Dataset creators should read through these questions prior to any data collection and then provide answers once data collection is complete. Most of the questions here are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for their chosen tasks. Some of the questions are designed to elicit information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.

	Question	Answer
1	**What do the instances that comprise the dataset represent (for example, documents, photos, people, and countries)?** Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.	The instances represent the date of the given instance, corresponding average sentiment and corresponding closing price. For each instance, there is one unique index, only one date, only one price and only one sentiment. There is only one type of instance.
2	**How many instances are there in total (of each type, if appropriate)?**	Not applicable.

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 (for example, 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 (for example, to cover a more diverse range of instances, because instances were withheld or unavailable	It contains all possible combinations of sentiment and price from Jan 2015 to Nov 2022. For every day, there are at least one news, so one sentiment score and price. We wanted to capture every possible point.
4	**What data does each instance consist of?** "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.	Each instance consists of one unique index, only one date, only one price and only one sentiment.
5	**Is there a label or target associated with each instance?** If so, please provide a description.	Yes, there are labels called “Index”, “Date”, “Price” and “Compound” for each instance.
6	**Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.	There is no missing data in our news dataset. However, as Brent Crude prices are not available during weekends, and there is no news on weekends, so we do not have any data on weekends.
7	**Are relationships between individual instances made explicit (for example, users' movie ratings, and social network links)?** If so, please describe how these relationships are made explicit.	Yes, the relationship between prices and news sentiment are made explicit via joining two tables with their “Date” columns.
8	**Are there recommended data splits (for example, training, development/validation, testing)?** If so, please describe these splits, explaining the rationale behind them.	We plan to do 80/20 split for training and testing and will use 10-fold cross validation. We will use it because it is a widely used one in the literature.
9	**Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.	There are no known errors or noises in the dataset.
10	**Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, and 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 (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.	This data-set is self-contained and contains information from two dataset that rely on external resources. You may refer to News Data and Prices Data to get information about them.
11	**Does the dataset contain data that might be considered confidential (for example, 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.	No, all of the information stored in the dataset can be published publicly.
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.	No. It only contains price and sentiment data which cannot be offensive.

If the dataset does not relate to people, you may skip the remaining questions in this section.

	Question	Answer
1	**Does the dataset identify any sub-populations (for example, by age, or gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.**	Not applicable.
2	**Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.**	Not applicable.

3	**Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race 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.**	Not applicable.
4	**Any other comments?**	-

Collection Process

As with the questions in the previous section, dataset creators should read through these questions prior to any data collection to flag potential issues and then provide answers once the collection is complete. In addition to the goals outlined earlier, the following questions are designed to elicit information that may help researchers and practitioners to create alternative datasets with similar characteristics. Again, questions that apply only to datasets that relate to people are grouped at the end of the section.

	Question	Answer
1	**How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.**	Data was directly observable. The News Dataset and the Prices Dataset were directly available through our datasets.
2	**What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, and software APIs)? How were these mechanisms or procedures validated?**	This dataset is created based on two excel files and uses Python to do so.
3	**If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?**	Dataset uses all available data.
4	**Who was involved in the data collection process (for example, students, crowd workers, contractors) and how were they compensated (for example, how much were crowd workers paid)?**	Dataset is collected by the WhiteRabbit team. No additional compensation was given, and the process was considered as their regular working hours.
5	**Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.**	Data is from Jan 2015 to November 2022. As the data continuously comes, the timeframe always matches.
6	**Were any ethical review processes conducted (for example, by an institutional review board)? If so, please describe these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.**	No ethical review is conducted. This is because of two reasons: 1. We do not have any sensitive or private information in the dataset. 2. The website indicates that all of its data is publicly available.

If the dataset does not relate to people, you may skip the remaining questions in this section.

	Question	Answer
1	**Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?**	The data is obtained via third parties.
2	**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**	Not applicable.

3	**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.**	Not applicable.
4	**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 applicable.
5	**Has an analysis of the potential impact of the dataset and its use on data subjects (for example, 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 applicable.
6	**Any other comments?**	-

Preprocessing/ Cleaning/ Labeling

Dataset creators should read through these questions prior to any preprocessing, 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.

Question	Answer
Was any preprocessing/cleaning/labeling of the data done (for example, 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 remaining questions in this section.	Yes. For the sentiment data, we used News Data and extracted a sentiment score by using NLTK's Vader functionality. It does all necessary preprocessing for sentiment scoring.
Was the "raw" data saved in addition to the preprocessed /cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.	In the creation process, there are some intermediate data frames created that have raw data as well. However, they are not saved as a different column as we do not think they will be needed at anytime.
Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.	We used https://bitbucket.org/whiterabbitt/data/src/main/final_data.py in our Bitbucket for this purpose.
Any other comments?	-

Uses

The following 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.

Question	Answer
Has the dataset been used for any tasks already? If so, please provide a description.	This dataset is newly created. Not applicable.
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.
What (other) tasks could the dataset be used for?	Apart from Brent Crude price forecasting, this dataset can be used to train sentiment classifiers for financial purposes. However, for this purpose, manual labeling is needed.

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 dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?	This dataset is a combination of Price Data and News Data. Problems are discussed in their datasheets.
Are there tasks for which the dataset should not be used? If so, please provide a description.	The dataset should only be used on Oil Prices. Everything else should be avoided.
Any other comments?	-

Distribution

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

Question	Answer
Will the dataset be distributed to third parties outside of the entity (for example, a company, institution, or organization) on behalf of which the dataset was created? If so, please provide a description.	Yes, the dataset will be publicly available for any kind of user as well as users of the Rabbidity Project.
How will the dataset be distributed (for example, tarball on the website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?	The dataset link will be available on Bitbucket.
When will the dataset be distributed?	The dataset will be distributed when the Rabbidity website is released (20.12.2022).
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.	We will use the Creative Commons Zero (CC0) license to openly share their datasets worldwide.
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.
Do any export controls or other regulatory restrictions apply to the dataset or individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.	No, there will not be any export controls or other regulatory restrictions.
Any other comments?	-

Maintenance

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

Question	Answer
Who will be supporting/hosting/maintaining the dataset?	@ Mert Mecit will be responsible for the dataset-related operations.
How can the owner/curator/manager of the dataset be contacted (for example, email address)?	He can be contacted via email or phone number available on Rabbidity Website.
Is there an erratum? If so, please provide a link or other access point.	No.

Will the dataset be updated (for example, to correct labeling errors, add new instances, or delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub).	Yes, the dataset will be updated daily for new news data. Also, we will do regular checking on our datasets if there are any errors. The updates will be communicated via Bitbucket.
If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period and then deleted)? If so, please describe these limits and explain how they will be enforced.	Not applicable.
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 dataset consumers.	Yes, the older dataset will be hosted. They will be kept in Bitbucket. The obsolescence will be communicated via Bitbucket.
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 dataset consumers? If so, please provide a description.	They can do augmentation or contributions on their local platform. No one will be permitted to modify the original dataset.
Any other comments?	-