Datasheet for dataset "BeanCounter"

Questions from the Datasheets for Datasets paper, v7.

Jump to section:

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.

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.

BeanCounter is one of the largest business-oriented text dataset and is created to facilitate research in business domain NLP and toxicity in NLP datasets.

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

The BeanCounter dataset is created by Bradford Levy and Siyan Wang at University of Chicago Booth School of Business.

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.

There are no specific grants that supported the creation of the dataset; we acknowledge general financial support from University of Chicago Booth School of Business.

Any other comments?

No.

Composition

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.*

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.

The instances are publicly available financial disclosure textual documents filed on Securities and Exchange Comission's Electronic Data Gathering and Retrieval system (SEC EDGAR) by entities subject to the Securities Acts of 1933 and 1934, the Trust Indenture Act of 1939, and the Investment Company Act of 1940.

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

We collected 16,486,145 documents (instances) from more than 16,000 entities.

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).

We filter out documents containing very little text or high proportion of white space; see Appendix A in Wang and Levy (2024) for more details. We provide 3 configurations of the dataset: BeanCounter.clean, BeanCounter.final and BeanCounter.sample. BeanCounter.clean is the final set of documents that has been filtered out with the cleaning technique described in Appendix A.3. BeanCounter.final is the set of documents that have been deduplicated on document basis (see Appendix A.4) and BeanCounter.sample is a 1% random sample of the dataset stratified by year.

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:

- accession number: unique number assigned to each filing according to the entity's CIK, filing year and number of business days.
- file name: name of the document submission including the extension (e.g. .html or .txt).
- text: textual content of the document.
- filing type: indicated type of submission to fulfill a specific SEC regulation; more specific than form type; e.g. DEF 14A (filing type) vs. DEF (form type).
- attachment type: purpose of the document in the particular filing. The two main types are the main filing or exhibits (supplementary materials to the main filing).
- date: date of filing submission.
- form type: indicated type of submission to fulfill a particular SEC regulation (similar to filing type but less specific).

- the accepted timestamp: second-precise timestamp of when the document is accepted into SEC EDGAR.

Is there a label or target associated with each instance?

If so, please provide a description.

No.

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.

No information should be missing from instances.

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.

Instances are attachments to a particular filing, and each filing can contain one or more attachments. If the filing has more than one attachment (or instance), each attachment in the filing shares the same accession (i.e. the instances are linked by accession).

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.

The training set contains all data extracted from SEC's EDGAR betwen 1996-2023. The validation set contains 100MB (uncompressed) of documents sampled from the start of 2024 through end of February, 2024. The training and validation sets are partitioned by time to ensure that data in the validation set is largely new and unobserved in the training set, since most entities are required to file updated reports at least annually.

Are there any errors, sources of noise, or redundancies in the dataset?

If so, please provide a description.

Since the entities are responsible for producing the documents, there is a possibility of misreporting numbers or information in their filings. If these errors are found by the SEC, they can ask for corrections from these entities; otherwise, the errors can go undetected. For discussion on reducing redundancies in the dataset, please see Appendix A.3 and A.4 in the manuscript for details.

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.*

The dataset is self contained.

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.

No, the data does not contain any confidential information. All financial disclosures filed on SEC EDGAR is publicly available. Discussion regarding the license of SEC EDGAR data can be found in beginning of Section 3 in Wang and Levy (2024).

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

If so, please describe why.

We have conducted extensive toxicity analysis of the dataset and determined that it is lower in toxicity compared to other web-based datasets; details can be found in Section 3.4 of the manuscript. Discussions regarding the difference between BeanCounter and other web-based datasets can also be found in the conclusion.

Based manual inspection of toxic content in the dataset, we have found rare instances of toxic sentences in filings that include earnings call transcript or discussions of discriminatory communication (with examples) in the context of Human Resources training manuals.

Does the dataset relate to people?

If not, you may skip the remaining questions in this section.

A small portion of our dataset may related to people in so much as they are mentioned by the entities in our dataset. For example, Tim Cook may be mentioned in our data if Apple, or their competitors, discusses him.

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.

BeanCounter includes references of various subpopulations; we explicitly study the toxicity of text surrounding these mentions and details can be found in Section 3.3 and 3.4 of Wang and Levy (2024).

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.

The dataset can contain personally identifiable information; however, the entities have consented to making this information available. See beginning of Section 3 in manuscript for more detailed discussion.

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.

Any other comments?

No.

Collection process

[T]he answers to questions here may provide information that allow others to reconstruct the dataset without access to it.

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 dataset associated with each instance is derived from the SEC's daily archives of filings accepted by the EDGAR system. The EDGAR system accepts a variety of file formats. We process all text and HTML-based files to extracted formatted long-form text from each filing. Full details of the dataset construction process can be found in Appendix A of Wang and Levy (2024).

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?

The SEC publishes daily archives of all filings accepted by the EDGAR system. We downloaded these in an automated manner, retrying any failed downloads until they succeeded.

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

We process all text and HTML-based filings. The "sample" configuration of the BeanCounter dataset consists of a random sample of 1% of the full BeanCounter dataset. We sample this data stratified by year to ensure an even volume of tokens for each year.

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 authors completed all data collection activities themselves.

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.

The data was collected in February 2024 however the SEC EDGAR system is similar to an append only database where each filing is associated with a timestamp denoting the date and time it was accepted by EDGAR. In that sense, any data collected retroactively, e.g., a filing from 2014, is representative of its content at the time EDGAR accepted it.

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.

No.

Does the dataset relate to people?

If not, you may skip the remainder of the questions in this section.

A small portion of our dataset may related to people in so much as they are mentioned by the entities in our dataset. For example, Tim Cook may be mentioned in our data if Apple, or their competitors, discusses him.

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

All data is collected from SEC EDGAR.

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.

They were not.

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, all EDGAR filers consent to the SEC's terms of use, which stipulate that "Information presented on www.sec.gov is considered public information and may be copied or further distributed by users of the web site without the SEC’s permission." More details on the SEC's policy can be found here.

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.

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.

See Wang and Levy (2024) for a discussion of the implications and impact of the dataset.

Any other comments?

Preprocessing/cleaning/labeling

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.

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, filings which are both raw text and HTML-based had some preprocessing and cleaning applied. The goal of these steps is to extract long-form text from the original filings while preserving meaningful formatting such as paragraphs breaks, indentation, and lists. See Wang and Levy (2024) for further details of the exact preprocessing and cleaning.

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 is directly available from the SEC and they have pledged to continue to make it available.

Is the software used to preprocess/clean/label the instances available?

If so, please provide a link or other access point.

Yes, please see supplementary materials document for accessing it.

Any other comments?

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.

Has the dataset been used for any tasks already?

If so, please provide a description.

We explored the utility of BeanCounter by continually pretraining existing models on the dataset and evaluating it on financial and toxicity related tasks; see Section 4 of Wang and Levy (2024) for detailed discussion.

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, BeanCounter has not been used in other papers and systems.

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

The dataset could be used for tasks that evaluate social biases (e.g. CrowS-Pairs),truthfulness (e.g. TruthfulQA), timeliness (e.g. TempLAMA) and other financial domain knowledge evaluations (e.g. ConvFinQA).

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?

While we process all of the filings uploaded to EDGAR, our text extraction process only supports text and HTML-based documents. As a result, the content of other document types, e.g., Excel, will not appear in our dataset.

Are there tasks for which the dataset should not be used?

If so, please provide a description.

Due to the nature of content in the dataset, models trained on BeanCounter may lack imagination and perform poorly on benchmarks that evaluate the model's creativity; see Conclusion in Wang and Levy (2024) for additional discussions on the idiosyncracy of the data.

Any other comments?

No.

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.

Yes.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?

Does the dataset have a digital object identifier (DOI)?

The dataset will be available via HuggingFace Hub as a collection of gzipped json files.

When will the dataset be distributed?

It will be made publicly available close to the NeurIPS conference date.

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.

Yes, the dataset will be distributed under Open Data Commons Attributions license. This permissive license allows users to share and adapt the dataset as long as they give credit to the authors.

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 to individual instances?

If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

These questions are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan with dataset consumers.

Who is supporting/hosting/maintaining the dataset?

Bradford Levy and Siyan Wang are supporting and maintaining the dataset.

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

Please refer to the manuscript for email addresses.

Is there an erratum?

If so, please provide a link or other access point.

Please see the github repository for erratum related to the dataset.

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)?

Yes, as soon as practicable. The updates can be seen on Github and HuggingFace Hub.

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, the entities in the dataset have agreed to make it publicly available in perpetuity.

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.

Yes, the older versions of the dataset will continue to be hosted on Huggingface Hub.

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.

Researchers can interact and use the BeanCounter dataset via Huggingface Hub; we do not provide functionalities beyond what Huggingface Hub provides.

Any other comments?

No.