
Supplementary Materials for *BeanCounter: A low-toxicity, large-scale, and open dataset of business-oriented text*

Siyan Wang
University of Chicago
Chicago, IL 60637

Bradford Levy*
University of Chicago
Chicago, IL 60637

1 Dataset Documentation

We chose to use the “Datasheets for Datasets” framework to document our dataset. We have included that as a separate PDF “BeanCounter_Datasheet.pdf” and incorporated it into the page where our dataset will be hosted.

2 Accessing the Dataset

We have made the dataset available to the reviewers via Hugging Face Hub. We provide three configurations of the data:

1. “clean” which is the 159B tokens of cleaned non-deduplicated data
2. “final” which is the 111B tokens of cleaned and deduplicated data
3. “sample” which is a sample of roughly 1% of the data in “final” where we sample from yearly strata to ensure an even distribution of data by year

Since the dataset is not yet public, reviewers must use the provided token to download the data. The token and example code to download the data is provided below. The code assumes that Hugging Face Datasets has been installed, e.g., by running `pip install datasets`.

```
from datasets import load_dataset

token = "hf_ZJaDhLjjYYPEOfpmxhJdeoRR0fhmqYYvWS"

beancounter = load_dataset(
    "blevy41/BeanCounter",
    name="sample", # Load random sample, clean, or final
    token=token,
)
```

3 Croissant Metadata

We have included Croissant metadata in the form of a JSON-LD file “beancounter_croissant.json”. The Croissant metadata can also be access via Hugging Face’s API:

```
import json
import requests
```

*Corresponding author: bradford.levy@chicagobooth.edu

```
token = "hf_gnmxlkxGlmKdfLxkmjqebdNqoVByUr0Xxv"
headers = {"Authorization": f"Bearer {token}"}
dataset = "blevy41/BeanCounter"
url = f"https://huggingface.co/api/datasets/{dataset}/croissant"

response = requests.get(url, headers=headers)
print(json.dumps(response.json(), indent=2))
```

4 Author Statement

The authors acknowledge that they bear all responsibility in case of violation of rights, etc., and have released the dataset under the ODC-By license.

5 Hosting, Licensing, and Maintenance Plan

The authors plan to make the dataset publicly available via the Hugging Face Hub platform. The dataset will be released under the ODC-By license. The authors will accept maintenance requests via the project's GitHub repository and address them as needed while welcoming community engagement.

6 Accessing the Code

The authors also plan to make all code associated with the analyses in the paper publicly available. This will be done via the project's GitHub repository. Since the dataset is not yet public, we have made these files available via the DropBox link here. We have used DropBox because there are several gigabytes worth of data from intermediate analyses, e.g., cached responses from the Perspective API. If there is another way that the reviewers would prefer to access the data, please let the authors know and we will make that happen.