A Appendix

A.1 Dataset Details

This section provides additional details about the three datasets analyzed in this work. A repository with metadata and code for reproducing our results is available at `https://github.com/jackbandy/bookcorpus-datasheet`

A.1.1 Original BookCorpus Dataset

While BookCorpus is no longer publicly available, we obtained copies of the dataset files directly from the authors' website[5] where it was previously distributed [49]. Specifically, we obtained a directory called `books_txt_full` that contains 16 subdirectories corresponding to the genres of the books (e.g. Adventure, Fantasy, Historical, etc.), as well as a directory called `books_in_sentences` that contains two large files (`books_large_p1.txt` and `books_large_p2.txt`) with one sentence per line.

Within the `books_in_sentences` directory, the `books_large_p1.txt` file contained 536,517,284 words, and `books_large_p1.txt` contained 448,329,073 (based on the "-wc" unix word count software), for a total of 984,846,357 words. This aligns exactly with the number of words reported by [49] when introducing BookCorpus. Also, the first file contained 40,000,000 sentences, and the second file contained 34,004,228, together accounting for the 74,004,228 sentences reported by [49] when introducing BookCorpus.

The `books_txt_full` directory contained 11,040 text files, even though [49] report 11,038 books in the original BookCorpus dataset. Two extra files account for the discrepancy: one called `romance-all.txt` (1.12 GB) and another called `adventure-all.txt` (150.9 MB), large files which appear to contain concatenated text from all books within the respective genres.

However, the individual text files from the `books_txt_full` directory only contained 811,601,031 total words (after removing the `romance-all.txt` and `adventure-all.txt` files) – more than 170M words shy of the full sentence corpus. This is likely due in part to some empty files in the data we obtained (e.g. `Magic_of_the_Moonlight.txt` and `Song-of-Susannah.txt`), although we have no way to identify the complete causes of the word count discrepancy.

A.1.2 BookCorpusOpen Dataset

For the purposes of comparison, we also downloaded a newer, replicated version of BookCorpus which we refer to as *BookCorpusOpen*, in line with a publicly-maintained version of the dataset [16]. BookCorpusOpen is included in the Pile dataset as BookCorpus2 [19] and has been referred to by various other names (e.g. BookCorpusNew, Books1, OpenBookCorpus). The files we inspect include a list of 18,060 URLs from smashwords.com, and corresponding text files for 17,868 of them.

A.1.3 Smashwords21 "Superset"

To help address questions related to sampling, we collected a superset of the books represented in BookCorpus and BookCorpusOpen. Originally, BookCorpus contained all free English books from Smashwords which were longer than 20,000 words. A "complete" superset might contain all books ever published, or a similarly vast collection. For the purposes of this paper, we collected metadata about 411,826 unique books publicly listed on Smashwords as of April 2021.

To create this superset, we scraped all books listed on Smashwords, similar to what has been done in efforts to replicate BookCorpus [25]. Notably, our scrape recovered 411,826 unique books, while Smashwords reported that over 565,000 total books had been published on the website at the time. This discrepancy likely stems from a default filter that excludes adult erotica from the public listings. We could have set a `no_filtering` cookie in our scraping program to include these books, however, BookCorpus and BookCorpusOpen do not mention this, so we only scraped books that were publicly-listed by default. We ran the scraping program twice to help ensure coverage.

[5] We have notified the authors of the security vulnerability that allowed us to download the dataset.

Copies	Number of Books
1	4,255
2	2,101
3	741
4	82
5	6

Table 2: Number of unique books with different numbers of copies in BookCorpus. 4,255 books only had one copy in BookCorpus (i.e. not duplicated), 2,101 had two copies, etc.

Smashwords21 only contains book metadata (not the books themselves), scraped from the public web. As such, we make the data available through the MIT license, accessible at `https://github.com/jackbandy/bookcorpus-datasheet`

A.2 Duplicate Analysis

To analyze duplicate books in BookCorpus, we started by identifying potential duplicates based on file names. This step suggested that 2,930 books may be duplicated. Using the `diff` Unix program, we confirmed that BookCorpus contained duplicate, identical text files for all but five of these books, and manually inspected the five exceptions:

- `299560.txt` (Third Eye Patch), for which slightly different versions appeared in the "Thriller" and "Science Fiction" genre directories (only 30 lines differed)
- `529220.txt` (On the Rocks), for which slightly different versions appeared in the "Literature" and "Science Fiction" genre directories (only the title format differed)
- `Hopeless-1.txt`, for which identical versions appeared in the "New Adult" and "Young Adult" genre directories, and a truncated version appeared in the "Romance" directory (containing 30% of the full word count)
- `u4622.txt`, for which identical versions appeared in the "Romance" and "Young Adult" genre directories, and a slightly different version appeared in the "Science Fiction" directory (only 15 added lines)
- `u4899.txt`, for which a full version appeared in the "Young Adult" directory and a truncated version (containing the first 28 words) appeared in the "Science Fiction" directory

Combined with the `diff` results, our manual inspection confirmed that each filename represents one unique book, thus BookCorpus contains at most 7,185 unique books.

A.3 Supporting Tables

Table 2, Table 3, and Table 4 include full descriptive statistics for some standalone statisitcs referenced throughout the datasheet.

	BookCorpus	BookCorpusOpen	Smashwords21
Romance	26.1% (2880)	18.0% (3314)	16.0% (66083)
Fantasy	13.6% (1502)	17.2% (3171)	10.6% (44032)
Science Fiction	7.5% (823)	13.3% (2453)	7.8% (32063)
New Adult	6.9% (766)	0.9% (175)	0.7% (2902)
Young Adult	6.8% (748)	9.5% (1748)	4.6% (19015)
Thriller	5.9% (646)	7.4% (1368)	5.7% (23587)
Mystery	5.6% (621)	5.3% (987)	4.7% (19351)
Vampires	5.4% (600)	0.0% (0)	0.0% (0)
Horror	4.1% (448)	3.9% (727)	3.9% (15944)
Teen	3.9% (430)	9.5% (1752)	4.6% (19154)
Adventure	3.5% (390)	11.5% (2117)	7.1% (29474)
Other	3.3% (360)	0.1% (18)	0.3% (1075)
Literature	3.0% (330)	3.0% (560)	2.6% (10592)
Humor	2.4% (265)	4.1% (749)	3.0% (12333)
Historical	1.6% (178)	4.7% (864)	4.5% (18815)
Themes	0.5% (51)	1.3% (243)	1.5% (6179)

Table 3: Distribution of genres in the BookCorpus sample, compared to books in the new BookCorpusOpen dataset and all books listed on Smashwords as of April 2021. Smashwords21 does not contain duplicates (based on book URLs), though BookCorpus and BookCorpusOpen do contain duplicates.

	BookCorpusOpen	Smashwords21
Sikhism	0	15
Judaism	18	371
Islam	229	1305
Hinduism	12	261
Christianity	154	2671
Buddhism	32	512
Atheism	18	175

Table 4: Religious subject tally, for books with religious metadata in BookCorpusOpen (N=18,451) and Smashwords21 (N=411,826). Overall, Smashwords over-represents books about Christianity, though books about Islam are over-represented in the BookCorpusOpen sample.

A.4 Summary Data Card for BookCorpus

Dataset Facts

Dataset BookCorpus
Instances Per Dataset 7,185 unique books, 11,038 total

Motivation	
Original Authors	Zhu and Kiros et al. (2015) [49]
Original Use Case	Sentence embedding
Funding	Google, Samsung, NSERC, CIFAR, ONR

Composition	
Sample or Complete	Sample, ≈2% of smashwords.com in 2014
Missing Data	98 empty files, ≤655 truncated files
Sensitive Information	Author email addresses

Collection	
Sampling Strategy	Free books with ≥20,000 words
Ethical Review	None stated
Author Consent	None

Cleaning and Labeling	
Cleaning Done	None stated, some implicit
Labeling Done	None stated, genres by smashwords.com

Uses and Distribution	
Notable Uses	Language models (e.g. GPT [38], BERT [11])
Other Uses	List available on HuggingFace [15]
Original Distribution	Author website (now defunct) [49]
Replicate Distribution	BookCorpusOpen [16]

Maintenance and Evolution	
Corrections or Erratum	None
Methods to Extend	"Homemade BookCorpus" [25]
Replicate Maintainers	Shawn Presser [15]

Genres	% of BookCorpus*
Romance 2,881 books	26.1%
Fantasy 1,502 books	13.6%
Science Fiction 823 books	7.5%
Vampires 600 books	5.4%

Horror 4.1%
Adventure 3.5%
Historical Fiction 1.6%
• Teen 3.9%
• Literature 3.0%

Not a significant source of nonfiction.

* Percentages based on directories in books_txt_full. Some books cross-listed.