Journal of Open
Humanities Data

HATHI 1M: Introducing a Million Page Historical Prose Dataset in English from the Hathi Trust

RESEARCH PAPER

SUNYAM BAGGA

ANDREW PIPER

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

We present a new dataset built on prior work consisting of 1,671,370 randomly sampled pages of English-language prose roughly divided between modes of fictional and non-fictional writing and published between the years 1800 and 2000. In addition to focusing on the "page" as the basic bibliographic unit, our work employs a single predictive model for the historical period under consideration in contrast to prior work. Besides publication metadata, we also provide an enriched feature set of 107 features including part-of-speech tags, sentiment scores, word supersenses and more. Our data is designed to give researchers in the digital humanities large yet portable random samples of historical writing across two foundational modes of English prose writing. We present initial insights into transformations of linguistic patterns across this historical period using our enriched features as possible pointers to future work. The data can be accessed at https://doi.org/10.7910/DVN/HAKKUA.

CORRESPONDING AUTHOR:

Sunyam Bagga

txtLAB, McGill University, Montreal, Canada

sunyam.bagga@mail.mcgill.ca

KEYWORDS:

digital humanities; cultural heritage; English-language prose; fiction; predictive modeling

TO CITE THIS ARTICLE:

Bagga, S., & Piper, A. (2022). HATHI 1M: Introducing a Million Page Historical Prose Dataset in English from the Hathi Trust. *Journal of Open Humanities Data*, 8: 7, pp. 1–10. DOI: https://doi.org/10.5334/johd.71

1 CONTEXT AND MOTIVATION

Understanding the historical evolution of language and culture is a central mission of the humanities. Doing so can teach us about the contingency of ideas (versus their naturalization), the nature of intellectual, linguistic or creative influence, or the relationship between larger social forces and human creative behavior. The digitization of large historical collections poses a profound research opportunity for better understanding the nature of cultural change. However, sampling documents from the past poses significant challenges (Bode, 2020). In addition to the inevitable biases introduced through historical archiving practices, up to and including digitization, much of what is contained in historical archives is not richly indexed, providing limited knowledge of the specific nature of digital collections. Recent work has begun to prioritize this process of enriching historical collections, using emerging techniques in machine learning to identify, for example, genre-level (Underwood, Kimutis, & Witte, 2020) or visual-level (Fyfe & Ge, 2018; Piper, Wellmon, & Cheriet, 2020) qualities of digitized texts to support further research.

At the same time, major technical and legal hurdles remain in place inhibiting wider access to historical cultural data. Copyright restrictions continue to limit access to large amounts of material covering broad stretches of time. While admirable technical solutions have been implemented to make copyright-restricted material accessible, these solutions require high levels of technical expertise that exceed most researchers.[1]

It is with these challenges and opportunities in mind that we build on prior work (Underwood et al., 2020) to generate a large sample of richly annotated prose data in English drawn from the Hathi Trust Digital Library. The Hathi Trust Digital Library represents the largest collection of digitized historical documents in English, with over 17 million volumes. It therefore represents an ideal resource for the historical study of cultural documents up to the very recent past. Due to the aforementioned copyright restrictions, our data takes the form of page-level metadata, which include document IDs, estimated publication dates, author and titles, word and part-of-speech frequencies, and a small set of enriched higher-level features which we describe below in greater detail.

This data allows researchers to work directly with our corpus on questions of historical interest, whether focusing on specific linguistic or stylistic changes over time, larger thematic shifts by mode of writing, or even towards further refining the classificatory annotation we have undertaken here to develop finer-grained subgroupings within our data. Finally, researchers can also use the page IDs to refer back to the Hathi Trust collections to work directly with the full text data using their capsule system.

The procedures we use to annotate our prose data, which we describe more fully below, offer the following affordances for historical study.

The page as historical unit. Recent work has emphasized the importance of the "page" as a distinct historical artifact for both reading and viewing (Mak, 2011). Focusing on the page as the primary sampling unit of writing from the past can help address two key research challenges. First, it can help avoid problems of over-/under-sampling from long/short works. Given that historical collections contain works of very high levels of variability when it comes to page length, sampling limited numbers of pages from individual works can contribute to the collection of a broader cross-section of writing and avoid problems of longer works being overweighted. Second, the page also offers insights into the material conditions of writing that have been of interest to historians of the book. By focusing on the page as the sampling unit, researchers can recover physical traces of reading and writing from the page structure and observe how these vary over time given a representative sample of writing. Nevertheless, we also retain book-level metadata so that researchers can also work at the work-level if interested.

Comparative framework. For the period under investigation, writing in prose is by far the dominant mode during these years. While prior work has developed models for the detection of prose fiction during this period (Underwood et al., 2020), we build models for the creation of symmetrical collections of fictional and non-fictional modes of writing. Our work is motivated

[1] *https://wiki.htrc.illinois.edu/display/COM/HTRC+Data+Capsule+Environment* (last accessed: February 1, 2022).

by theoretical frameworks grounded in theories of social differentiation (Luhmann, 1995), where the meaning and function of different modes of communication evolve in distinction to one another. Our dataset allows researchers to study the distinctive stylistic qualities of these two foundational modes of modern communication in relation to each other.

Unified time frame. While there have been numerous frameworks of periodization proposed by literary and intellectual historians, with important differences by national boundaries, there is strong theoretical consensus that the years around 1800 mark an important historical caesura for English-speaking communities in the UK and North America, often related to political, epistemological, and industrial revolutions (Foucault, 2013; Koselleck, 2004). Our data thus contributes to the further study of periodization by allowing researchers to better understand stylistic and thematic variations within a larger time-frame beginning in 1800 and continuing through the end of the twentieth century.

Single model. Our data is generated from a single predictive model for each mode of writing (fiction/non-fiction) across the entire sampled time-frame based on manually reviewed training data derived from prior work (Underwood et al., 2020). While there is still much work to be done regarding the implications of applying predictive models on collections spanning long historical time frames, as we show the use of a single model overcomes important anomalies introduced by the conjoining of multiple models from prior work (see ***Figure 4***). We report details about our models, our training data, and the estimated accuracy of our predictions in Section 2.

Enriched metadata. While full text remains the gold standard for studying historical collections, for many documents full text is either inaccessible due to intellectual property restrictions or only accessible through more complex technical means, which may exceed the technical capacities of many researchers. Additionally, in many cases researchers desire extracted properties, which may be expensive to compute in terms of either time and/or personnel. We thus provide metadata on all sampled pages so that researchers can access the original data in its full text form using the Hathi Trust capsule system. However, we also provide a set of enriched features about our data derived using standardized methods in text analysis, which we describe more fully in Section 3.2. These features range from unigrams (which are already publicly accessible) to part-of-speech tags, sentence length, word "supersenses" and more. By downloading our data, researchers can begin to analyze large collections of historical prose without any further computational processing of the original text files.

Portability. Researchers working in the space of digital humanities have emphasized the importance of values such as portability and accessibility to facilitate the necessary cross-disciplinary investigation at the heart of the field (Schmidt, 2018). Our data attempts to strike a balance between size and portability, by which we mean the ability of the data to move easily between researchers and to be accessible in straightforward ways. Studying long time scales necessarily requires large data collections as each time unit (year/decade) becomes sparser the less data one has. Even 100 documents per year from a 200 year time period means working with a dataset of 20,000 observations. Rather than sample from all possible documents in the English collection of the Hathi Trust Digital Library, we choose to cap our samples at 5,000 pages per year per mode. While this gives us just over 1.6 million observations overall, the total word count is reduced considerably by the short length of pages. Our data thus consists of 588 million words, an overall size of 6.9GB for the full text, and 1.7GB for metadata and enriched features. Our aim is to provide data objects that are easily manipulable with minimal computing power, in order to balance ease of circulation with historical representativeness.

Our hope in providing this data is to give researchers interested in historical change a robust, reliable, and easily accessible foundation with which to begin their research. In addition to describing our annotation and collection methods in greater detail below, we also provide an initial overview of some stylistic changes observed in our data using our enriched features as a means of prompting future work. We conclude with a discussion of how to leverage this data to generate future collections beyond English held by the Hathi Trust Digital Library in order to address growing concerns about the relative lack of attention in the computational humanities to multilingual research questions.

2 METHOD

Our work utilizes machine learning to automatically label random samples of English-language volumes held by Hathi Trust according to two modes of writing in prose: fiction and non-fiction. We then sub-sample pages from these volumes to achieve our desired threshold of pages (ca. 1 million per class). We adopt a supervised learning approach which requires labeled instances of training data as input. In this section, we explain (1) the training-data used, (2) design and validation of the two classifiers (one each for fiction and non-fiction), and (3) our methodology to construct our final datasets using the validated classifiers.

2.1 TRAINING DATA

For this project, we have two separate classification tasks involving the detection of two modes of prose writing: "fiction" and "non-fiction." Our two primary sets of training data thus contain the following manually curated data:

1. **Fiction (411 documents):** We randomly sample two or three volumes per year from the class of English-language fiction drawn from the "most frequently reproduced" fiction in Underwood et al. (2020). Our training data is thus designed to represent a yearly cross-section of the most reproduced English-language fiction between 1800 and 2000. Year of publication is estimated using the "inferred date" category of Underwood et al. (2020). We only sample from volumes that have a minimum of two or more reprints during our period. This process results in 411 volumes that were manually reviewed for appropriateness.
2. **Non-Fiction (400 documents):** We randomly sample two volumes per year from English-language volumes labeled as 'not fiction' from the HathiTrust Digital Library published between the years 1800 to 1999. We utilize the labels embedded in the MARC-XML of the volumes which was extracted using the HathiTrust Bibliographic API. These were then manually filtered and verified by the authors and the process repeated, resulting in a final collection of 400 non-fiction volumes spanning the years 1800-1999 with two volumes sampled per year.

It is important to note that while the classification tasks are similar, they are not symmetrical. The opposing class of "fiction," for example, is not simply "non-fiction," but also includes non-prose literary genres such as plays and poems (i.e., "not-fiction "). Similarly, the opposing class of "non-fiction" is not strictly "fiction" but also includes non-prose literary genres (i.e., "not-nonfiction"). In order to address this, we manually augment our opposing class of training data with non-prose literary documents as well as use heuristic rules to automatically remove these kinds of documents from our final annotated data. In this way, we ensure that our final data only includes prose data.

2.2 CLASSIFIER SELECTION

We use the Support Vector Machines (SVM) algorithm as our machine learning classifier, which has been shown to be among the best performing algorithms on a variety of text classification tasks (Joachims, 1998). A crucial aspect of the text classification pipeline is feature representation. We represent the input text (Hathi volumes) as bag of word n-grams which is one of the most simple yet effective methods for feature vectorization. In order to design an accurate classifier, we experiment with word unigrams, bigrams, trigrams and all three word n-grams combined, and pick the one which yields the best performance in a 10-fold cross validation test. Simultaneously, we perform cross-validated tuning for SVM's hyperparameters namely, *C* and kernel. In addition, we also optimize SVM's probability threshold for predicting fiction and the different n-grams' vocabulary sizes. Note that we filter out the paratext during the training phase of the classifier by only using the middle 60% text of the volume and removing running headers, footers and page numbers.

Validation Results. In the cross-validation setting within the 811 volumes, SVM achieves an average f1 score of 0.968 and 0.964 for the Fiction and the NonFiction classifier respectively. Both classifiers yield the best results when the feature space consists of a combination of word unigrams, bigrams and trigrams. All the optimal parameters for both classifiers are listed in *Table 1*.

CLASSIFIER	FEATURE-SPACE	PROB-THRESHOLD	SVM HYPERPARAMETERS
Fiction	top-1k word uni-, bi-, trigrams	0.75	*C* = 1 & Gaussian Kernel
Non-Fiction	top-100k word uni-, bi-, trigrams	0.70	*C* = 1 & Linear Kernel

Table 1 The optimal feature-space and hyperparameters obtained using 10-fold cross-validation for our SVM classifier.

2.3 VOLUME AND PAGE SAMPLING

After validating our learning model with the optimal feature space and hyperparameters derived via cross-validation, we next describe our pipeline for constructing each of the datasets. This involves a three-step process of sampling volumes from Hathi Trust; running our classifiers on the sampled volumes; and then sampling pages from the classified volumes.

Sampling volumes. For each year of our designated time period, we randomly sample volumes from the Hathi Trust collection until one of two conditions are met: either we run out of volumes to sample or we reach our threshold number of yearly pages. The publication date is accessed using the *rights_date_used* field from Hathifiles. While sampling, we use a heuristic set of title and genre keywords to remove inappropriate volumes from each of our classes prior to classification. For example, for Fiction and Non-Fiction we discard all non-prose literary volumes by conditioning on words such as 'poet', 'poem', 'a comedy', 'a tragedy', etc., that appear in either the title or genre labels. For Non-Fiction, we also remove dictionaries and cookbooks using a similar set of heuristics. The complete list of heuristics can be found in the Heuristics Section of the readme file released with the dataset. In addition to the keyword-based filtering, we also remove volumes with duplicate titles, i.e., we remove a volume if the Levenshtein Distance between its title and any title in the sampled-set is more than a threshold of 90 (manually validated). Thus for a given year we attempt to condition on unique works; however, we do not do this across years such that works that are multiply reprinted in subsequent years have the chance of being resampled. Our data is designed to condition on writing published in a given year rather than writing composed in a given year.

Sampling pages. After sampling volumes, we then run our classifiers on the sampled volumes to predict each volume's class. We run once for each target class (Fiction/Non-Fiction). From each of the newly classified volumes within both classes, we then sample up to five random pages that meet the following quality constraints. The page must:

1. contain more than 100 words and 2 sentences. Pages are processed through BookNLP (Bamman, Underwood, & Smith, 2014) for word-tokenization and sentence-tokenization.
2. be in English. The English-language check is completed using Google's language-detection library.
3. possess an OCR quality of 80%. As a proxy for OCR quality, we use the percentage of English words on that page. Therefore, any page that has less than 80% English words will not be included in the dataset. The true set of English words comes from our lexicon of ~20,000 words drawn from English-language novels published between 1800–2000.
4. appear in the middle 60% of the volume to avoid sampling paratext related to introductory and advertising material.

3 RESULTS

3.1 DATASET

The execution of the pipeline discussed in Section 2.3 produces a collection of 765,920 fiction pages and 905,450 non-fiction pages published between 1800 and 2000 that were sampled from 153,184 and 181,090 volumes respectively. ***Figure 1*** shows the distribution of those pages across the two centuries. As can be seen, we are not able to achieve the 5,000 page mark for fiction for most of the 19th century. This is because the HathiTrust Digital Library did not have enough volumes from the early 19th century that were classified as fiction by our learning model. Another smaller contributing factor to this deficit is the fact that there are many volumes in the HathiTrust collection for which the publication date is not available. This leads to the imbalance in the number of pages in each of the collections.

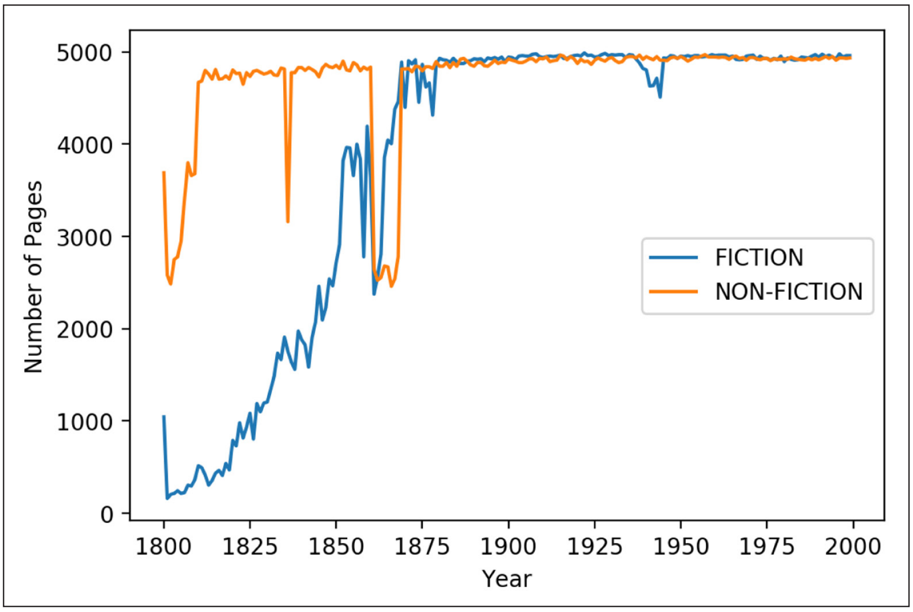

Figure 1 The distribution of the number of fiction and non-fiction pages across the two hundred years of our dataset.

HTID	YEAR	TITLE	AUTHOR	PAGE NUMBERS
nyp.33433090062153	1947	Rebel halfback	Archibald, Joe, 1898-	[50, 91, 96, 155, 159]
emu.010002632416	1852	The soldier of fortune	Curling, Henry, 1803–1864.	[159, 91, 204, 155, 166]
emu.010002426066	1886	Virginia the American	Edwardes, Charles.	[159, 204, 250, 166, 155]
emu.010002588974	1895	Moths/by Ouida	Ouida, 1839–1908.	[166, 155, 250, 204, 159]

Table 2 A preview of the columns and rows in the metadata table released for Hathi-Fiction.

We make publicly available the metadata and enriched features for both of our collections. Since most of the dataset is protected by copyright, we are not able to release the full text of pages. Metadata contains the following fields, for which we provide a small sample in ***Table 2***:

- Volume Identifier: This is a permanent and unique HathiTrust item identifier, referred to as *htid* in the dataset.
- Title: This refers to the title of the work, which in some cases may also include an author.
- Author: This refers to the name of the person, company or meeting that created the work.
- Year: This is the derived publication date of the item in the Hathifiles.
- Page Numbers: This is a list of five page numbers per volume that meet our quality constraints.

3.2 ENRICHED FEATURES

In order to provide researchers with deeper knowledge about the dataset and potentially applicable stylistic features for historical study, we compute and make publicly available a set of enriched features for both collections.

Before processing our enriched features, we engage in the following preparatory steps to remove "paratext" common to digitized printed objects. This includes the removal of running headers, chapter headings, footers, empty lines and other OCR-related noise. This is achieved by processing each page line by line and applying a set of hand-crafted rules to remove unwanted text.

After removal of paratext, we compute the following features for each page in our dataset:

1. *TotalLines*: The number of text lines per page.
2. *TotalWords*: The number of words (lexemes) per page. This does not include punctuation.

3. *TotalTokens*: The number of total tokens (including punctuation) per page.
4. *AvgSentlen*: Average sentence length is defined as the ratio of the total number of words to the total number of sentences per page.
5. *PctDialog*: Percent dialogue on the page is computed as the ratio of the number of words in quotes to the total number of words on the page. The *Quotation label* of BookNLP is used to determine whether a given word is inside quotation marks or not.
6. *Tuldava*: This is a common measure of readability that corrects for extreme scores that can result from metrics such as Flesch's Reading Ease and is more comparable across languages. It is computed using: $\frac{\#Syllables}{\#Words} \log \frac{\#Words}{\#Sentences}$
7. *Sentiment*: We calculate the sentiment score of each page using VADER, a popular lexicon and rule-based sentiment analysis tool (Hutto & Gilbert, 2014). It returns four scores: a normalized weighted composite score between –1 (most extreme negative) and +1 (most extreme positive); the other three are scores for 'pos', 'neg', 'neu' which sum up to 1. We include all four scores in our feature set.
8. *Emotion*: We use the NRC Word-Emotion Association Lexicon to calculate emotion scores per page (Mohammad & Turney, 2013). The lexicon consists of a list of English words and their associations with ten basic emotions and sentiment (anger, fear, anticipation, trust, surprise, sadness, joy, disgust, positive, negative). We include all 10 scores in our feature set. Note that each count is normalized by the total number of words on the page.
9. *Part-of-speech*: We process our pages through BookNLP and release the frequency of the part-of-speech tags for each page. The frequency is normalized by the total number of words on the page.
10. *Supersense*: Lastly, we include the frequency of BookNLP supersense tags on each page. The frequency is normalized by the total number of words.

In ***Figures 2*** and ***3***, we provide some sample graphs of individual features for both collections that point to future research opportunities. As we can see in ***Figure 2***, artifacts, which are defined as man-made objects according to the Wordnet taxonomy, are one of the fastest growing lexical features of both fiction and non-fiction. While the rise of artifacts levels off in non-fiction after 1950, it continues to rise in fiction suggesting that further exploration into this fictional investment in man-made objects is worth further study. Second, in ***Figure 3*** we see a major disconnect in the changing "difficulty" of fiction and non-fiction over the past two centuries. Where factual writing has remained relatively constant, if not gotten more difficult, fiction has steadily become easier to read from the perspective of word and sentence length, although a steady state consensus appears to have been achieved by 1950. The declining difficulty of fiction runs counter to scholarly narratives that have emphasized stylistic difficulty as a defining characteristic of literary modernism. Whether this is simply a matter of observing different kinds of writing or different stylistic emphasis, "readability" nonetheless marks another interesting area for further research in terms of fiction's identity over the course of the twentieth century as does fiction's change in "positivity" that we also see indicated in the bottom row of ***Figure 3***.

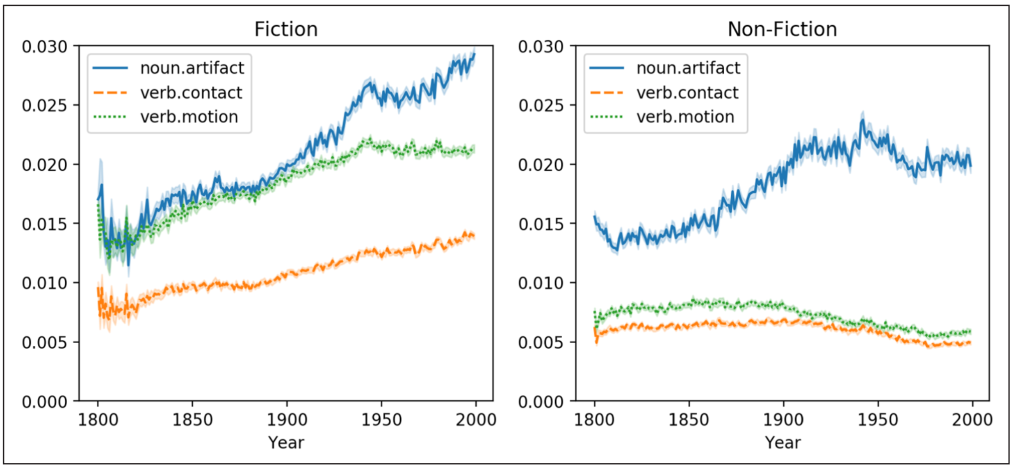

Figure 2 The distribution of three BookNLP supersense categories – artifacts, contact verbs, motion verbs – for pages sampled from 1800 to 1999. The left column corresponds to our fiction data and the right column is for our non-fiction data.

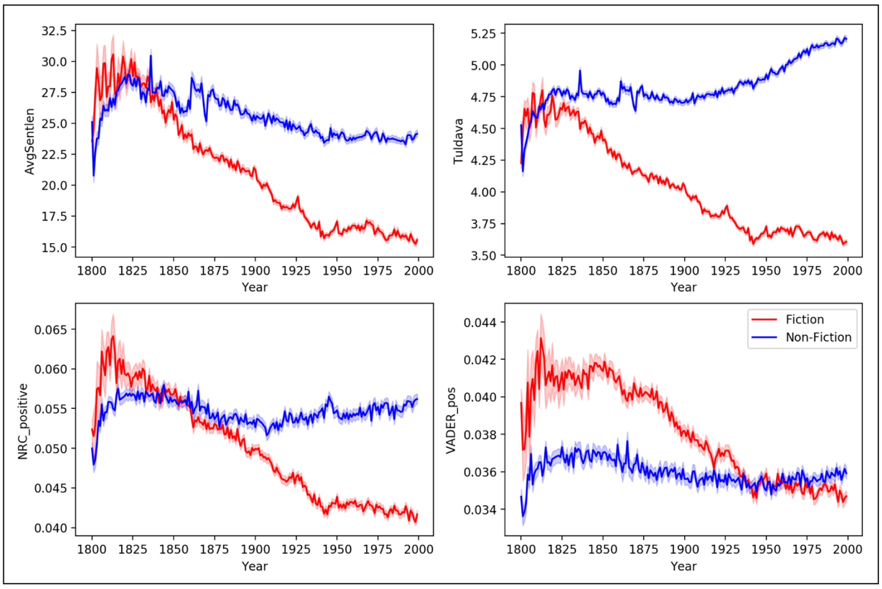

Figure 3 The distribution of four features from our Enriched Feature set – average sentence length, Tuldava score, NRC positive score, and VADER positive score – across our dataset of fiction pages (red) and non-fiction pages (blue) sampled from 1800 to 1999.

4 DISCUSSION

Our aim in constructing this dataset is to give researchers a portable yet extensive representative sample of historical prose in English across a two-hundred year time-span that covers two major modes of writing. With minimal technical requirements, researchers can begin to build on prior work using the Hathi Trust collections (Organisciak, Schmidt, & Downie, 2022; Schmidt, 2018; Underwood, 2019; Wilkens, 2021) and further explore comparative stylistic questions across our enriched features, as well as individual word features that are already publicly available through the Hathi Trust. In this section, we address the advantages and limitations of this dataset for historical study as well as identifying key avenues for future work.

The value of a single model. Underwood et al. (2020) were forerunners in producing the automated generation of genre labels within the Hathi Trust. Our work would not be possible without this prior work. However, as can be seen in ***Figure 4***, this dataset exhibits notable anomalies with respect to certain stylistic features that occur around specific historical junctures (notably the year 1900). As indicated in Underwood et al. (2020), these anomalies are the result of conjoining different predictive models and not underlying historical changes. To address this problem, we use a single predictive model across the entire historical period. We do not observe the same behavior in our dataset.

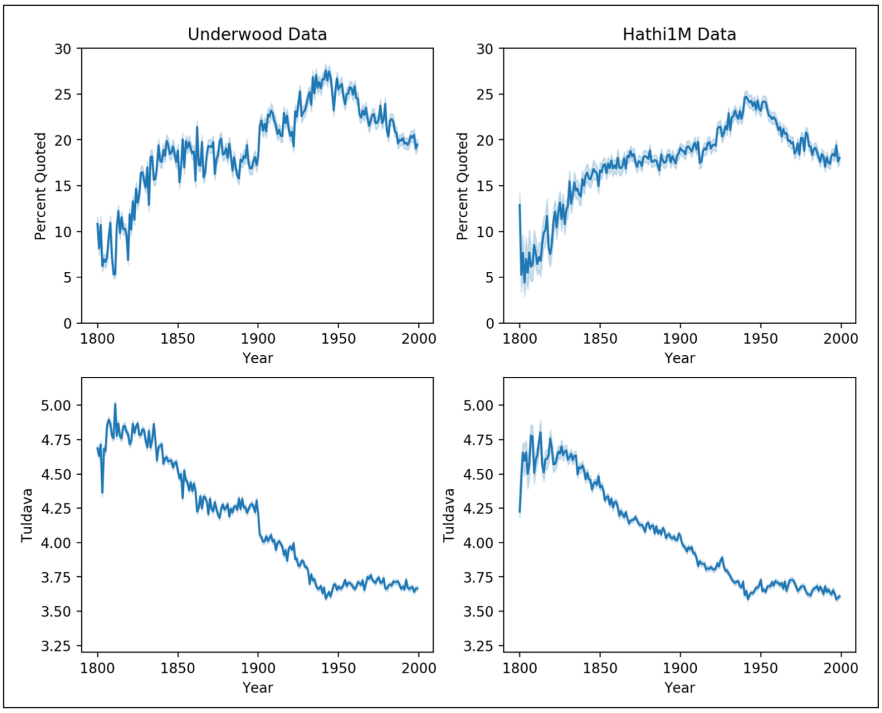

Figure 4 The distribution of % dialog and Tuldava scores for pages sampled from 1800 to 1999. The left column corresponds to the dataset derived from Underwood et al. (2020) and the right column corresponds to our Hathi1M fiction data.

The limitations of archives. An important limitation to highlight for future research is that while we have attempted to generate random samples of pages from within the Hathi Trust Digital Library, heritage collections and historical archives such as Hathi are not unbiased representations of the past. The documents contained within even seemingly massive digital collections are subject to curatorial and digitization pressures that potentially impact our understanding of the past. Future work will want to continue to understand, where possible, the relationship between the stylistic representations observed and their relationship to the biases introduced by different curatorial systems.

Beyond Fiction and Non-Fiction. While we provide a larger framework for comparative analysis across two major modes of writing, future work will want to concentrate on the annotation of further modes of writing. For example, non-fiction is an amalgam of a broadly heterogenous set of writing, including non-fictional narrative (memoirs and biographies), descriptive writing (travel writing), and argumentative or expository writing (scientific reports). Similarly, "fiction" incorporates many different kinds of genres that could be further specified in the data and that we assume exhibit meaningful stylistic differences.

Beyond English. Recent work has highlighted the need to move beyond largely anglophone representations of cultural practices using computational methods (Gil & Ortega, 2016). One major opportunity we see is the use of our data to further the detection of non-English collections within digital heritage collections. Recent advances in multi-lingual classification suggest that collections such as ours can serve as reliable training data for the detection of similar classes across numerous languages (Conneau et al., 2020). While such work requires computing resources that exceed current capacities, we look forward to finding future solutions to generate multi-lingual prose datasets for further cross-cultural research.

ADDITIONAL FILE

The files for metadata and the Enriched-Feature set for fiction and non-fiction have been uploaded to the Journal of Open Humanities Data Dataverse. It can be accessed at *https://doi.org/10.7910/DVN/HAKKUA*.

FUNDING STATEMENT

This work is supported by funding from the Social Sciences and Humanities Research Council of Canada (895-2013-1011).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Sunyam Bagga: Data Curation, Investigation, Writing; Andrew Piper: Conceptualization, Methodology, Writing.

AUTHOR AFFILIATIONS

Sunyam Bagga *orcid.org/0000-0001-6994-2192*
txtLAB, McGill University, Montreal, Canada
Andrew Piper *orcid.org/0000-0001-9663-5999*
txtLAB, McGill University, Montreal, Canada

REFERENCES

Bamman, D., Underwood, T., & **Smith, N. A.** (2014). A bayesian mixed effects model of literary character. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 370–379). DOI: *https://doi.org/10.3115/v1/P14-1035*

Bode, K. (2020). Why you can't model away bias. *Modern Language Quarterly, 81*(1), 95–124. DOI: *https://doi.org/10.1215/00267929-7933102*

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., et al. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8440–8451). Online: Association for Computational Linguistics. Retrieved from *https://aclanthology.org/2020.acl-main.747*. DOI: *https://doi.org/10.18653/v1/2020.acl-main.747*

Foucault, M. (2013). *Archaeology of knowledge*. London: Routledge. DOI: *https://doi.org/10.4324/9780203604168*

Fyfe, P., & **Ge, Q.** (2018). Image analytics and the nineteenth-century illustrated newspaper. *Journal of Cultural Analytics, 1*(2), 11032. DOI: *https://doi.org/10.22148/16.026*

Gil, A., & **Ortega, É.** (2016). Global outlooks in digital humanities: Multilingual practices and minimal computing. In *Doing digital humanities* (pp. 58–70). Routledge.

Hutto, C., & **Gilbert, E.** (2014). *Vader: A parsimonious rule-based model for sentiment analysis of social media text*. Retrieved from *https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109*

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning (ecml)* (pp. 137–142). Berlin: Springer. DOI: *https://doi.org/10.1007/BFb0026683*

Koselleck, R. (2004). *Futures past: on the semantics of historical time*. Columbia University Press.

Luhmann, N. (1995). *Social systems*. Stanford: Stanford University Press.

Mak, B. (2011). *How the page matters*. Toronto: University of Toronto Press.

Mohammad, S. M., & **Turney, P. D.** (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence, 29*(3), 436–465. DOI: *https://doi.org/10.1111/j.1467-8640.2012.00460.x*

Organisciak, P., Schmidt, B. M., & **Downie, J. S.** (2022). Giving shape to large digital libraries through exploratory data analysis. *Journal of the Association for Information Science and Technology, 73*(2), 317–332. Retrieved from *https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24547*. DOI: *https://doi.org/10.1002/asi.24547*

Piper, A., Wellmon, C., & **Cheriet, M.** (2020). The page image: Towards a visual history of digital documents. *Book History, 23*(1), 365–397. DOI: *https://doi.org/10.1353/bh.2020.0010*

Schmidt, B. (2018). Stable random projection: Lightweight, general-purpose dimensionality reduction for digitized libraries. *Journal of Cultural Analytics, 3*(1). DOI: *https://doi.org/10.22148/16.025*

Underwood, T. (2019). *Distant horizons: digital evidence and literary change*. Chicago: University of Chicago Press. DOI: *https://doi.org/10.7208/chicago/9780226612973.001.0001*

Underwood, T., Kimutis, P., & **Witte, J.** (2020). NovelTM datasets for english-language fiction, 1700–2009. *Journal of Cultural Analytics, 5*(2). DOI: *https://doi.org/10.22148/001c.13147*

Wilkens, M. (2021). Too isolated, too insular: American literature and the world. *Journal of Cultural Analytics, 6*(3). DOI: *https://doi.org/10.22148/001c.25273*

TO CITE THIS ARTICLE:
Bagga, S., & Piper, A. (2022). HATHI 1M: Introducing a Million Page Historical Prose Dataset in English from the Hathi Trust. *Journal of Open Humanities Data*, 8: 7, pp. 1–10. DOI: *https://doi.org/10.5334/johd.71*

Published: 11 March 2022

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.