# Piloting A Machine Learning Approach to Identify English-Language Fiction in the HathiTrust Digital Library

**Dubnicek, Ryan**
rdubnic2@illinois.edu
HathiTrust Research Center, Information Sciences, University of Illinois, United States of America

**Underwood, Ted**
tunder@illinois.edu
English and Information Sciences, University of Illinois, United States of America

## Motivation

As more text becomes open to text and data mining, a major challenge in finding and assembling a corpus of relevant items remains. In large digital libraries, such as the HathiTrust, comprising over 17 million items of varied format, language, and genre, metadata alone is not sufficient to identify items of interest. This is true for all volumes, where metadata records are often incomplete and particularly challenging for fiction volumes, where metadata standards are often too broad for specific analysis tasks (Miller 2000) even when present. This has led researchers to novel methods of classifying text, including analysis of stylometrics and textual features (Bucher 2018) and predictive modeling (Short 2019; Gupta 2019). This project will leverage the latter, and HathiTrust Research Center's Extracted Features Dataset (Jett et al. 2020), to build on successful classification efforts done as part of the NovelTM project (NovelTM 2022), improving the process and accuracy of Underwood, Kimutis and Witte's *NovelTM Datasets for English Language Fiction* (2020), while also seeking to expand the initial set of identified items by classifying volumes new to HathiTrust since the NovelTM dataset was first generated. This short paper will detail the methodology, early results, and planned future work in generating this dataset.

## Methods

Building on *NovelTM Datasets for English-Language Fiction*, this project used the HathiTrust Research Center's (HTRC's) Extracted Features (EF) Dataset to train a predictive model for English-language fiction. We differ from the NovelTM classification process by making predictions at the volume-level, using the tokens for each volume as input features for the model and metadata records included with the volumes as ground truth, supplemented by a more accurate manually-tagged subset of 2,730 volumes (Underwood et al. 2020). Three different statistical models were tested for the classification process: logistic regression (LR), support vector machine (SVM) and random forest (RF) using 120 trees, each implemented via the scikit-learn Python library (Pedregosa et al. 2011). To test the best model and process, we first assembled three samples of between 9,000-10,000 volumes each, gathered the HTRC EF data for each volume, and then split each set into 80% train and 20% test volumes. The Three samples are:

- Sample 1: 10,108 random volumes, matching distribution of items added to HTDL since 2016, by decade, yielding 1,605 fiction, 8,503 non-fiction.
- Sample 2: 9,969 random volume s, with the same selection logic as sample 1, but incorporating as many manually-verified fiction vols from NovelTM dataset as possible, yielding 1,580 fiction, 8,389 non-fiction volumes
- Sample 3: 9,061 volumes, including 53 F and 211 NF volumes for every decade represented in items added to HTDL since 2016, creating a train/test with equal volumes for each decade, yielding 1,328 fiction, 7,733 non-fiction volumes

After initial runs of each sample, we also benchmark a model that incorporates corrected ground truth for Sample 3, where about half of initial classification errors were incorrect F or NF classification. LR, SVM, and RF models were benchmarked via precision, recall, and F1 scores against each sample described above as well as the corrected Sample 3. The results for each model are in Table 1, with LR generally yielding the highest levels of accuracy, especially when applied to the corrected Sample 3. However, SVM performed well on Sample 3 - Corrected as well, and RF yielded similar levels of accuracy, with a lowered recall, for the uncorrected Sample 3.

Table 1. Precision, Recall, F1 scores and mean values for each sample and statistical model, logistic regression (LR), support vector machine (SVM) and random forest (RF). Bold indicates the highest value for each column and the highest mean value for P, R and F1 in the bottom row.

| | Logistic Regression | | | Support Vector Machine | | | Random Forest | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 |
| Sample 1 | 0.7838 | 0.9755 | 0.8692 | 0.8384 | 0.9205 | 0.8776 | 0.8665 | 0.8930 | 0.8795 |
| Sample 2 | 0.8589 | 0.9470 | 0.9008 | 0.8885 | 0.9238 | 0.9058 | 0.8824 | 0.8940 | 0.8882 |
| Sample 3 | 0.8804 | 0.9199 | 0.8997 | 0.9286 | 0.8750 | 0.9010 | 0.9697 | 0.8889 | 0.9275 |
| Sample 3 - Corrected | 0.9249 | 0.9506 | 0.9376 | 0.9702 | 0.9043 | 0.9361 | 0.9689 | 0.8642 | 0.9135 |
| Mean values | 0.8620 | 0.9483 | 0.9018 | 0.9064 | 0.9059 | 0.9051 | 0.9219 | 0.8850 | 0.9022 |

We also benchmarked mean F1 scores over five-fold cross-validation for each sample and model, the results of which are in Table 2 below. For each model, Sample 3 - Corrected produced the highest F1 score, a result to be predicted with a higher accuracy of ground truth. Sample 1 had the lowest scores over each model, with Sample 3 performing slightly better, but still behind Sample 2, which produced results only marginally worse than the corrected sample, which is a result that again speaks to both the value of reliable ground truth and the power of manually-verified training sets.

Table 2. Ranked mean F1 scores for each of four train/test sample datasets and three statistical models.

| | LR | SVM | RF | Rank |
|---|---|---|---|---|
| Sample 1 | 0.8815 | 0.8876 | 0.8744 | 3 |
| Sample 2 | 0.9123 | 0.9125 | 0.9111 | 2 |
| Sample 3 | 0.9023 | 0.8963 | 0.8989 | 4 |
| Sample 3 - Corrected | 0.9217 | 0.9180 | 0.9164 | 1 |

# Initial Results

Each of the models performed well by some metric, with best mean precision, recall and F1 scores belonging to unique models–RF yielding the highest mean precision (0.9219), LR yielding the highest mean recall (0.9483), and SVM with the highest mean F1 (0.9051), across all four samples. The best scoring models overall were Logistic Regression and SVM, with both scoring ~94% on F1, and LR and SVM performing better by five points on recall and precision, respectively. Seemingly either model would perform well in a larger classification task, and the deciding factor could be efficiency, which has not been explored for this study. Sample 3 - Corrected scored the best over LR and SVM, but findings may hint that a corrected Sample 2 could rival these results, as manual ground truth correction was worth an additional four points in accuracy for Sample 3. A four points improvement for Sample 2 would just edge out Sample 3 - Corrected for top accuracy.

Manual review and correction of errors for Sample 3 found four main types of errors:

- **Incorrect ground truth:** these are volumes incorrectly tagged as fiction or not fiction in their metadata. Examples of these volumes are as straight forward as Stephen Crane's *The Red Badge of Courage* or *Wuthering Heights* by Emily Bronte incorrectly being marked as not fiction.
- **Volumes that blur the lines of fiction, such as memoir, biography, or travel narrative:** volumes that look like typical fiction or not fiction, but are the inverse. These examples challenge our binary classification approach, and indeed would likely challenge a blind reviewer (but present many interesting research possibilities). Examples of these volumes are Daniel Defoe's *Robinson Crusoe* or John Hanning Speke's *Journal of the Discovery of the Source of the Nile*—the former being a "fake" travel narrative and the latter purporting to be authentic.
- **Non-prose fiction:** volumes that look like or are constructed from similar words as fiction, but are not prose, and thus not correct findings for our dataset. These include books of poetry and dramas.
- **True errors:** the last and least frequent errors are true errors—volumes the model just got plain wrong. Examples include annotated volumes, such as *The Works of Dr. Jonathan Swift*, Ward Greene's collection of prominent historical news stories, S *tar Reporters and 34 of Their Stories*, or a bound anthology of *Frank Leslie's Lady's Magazine.*

# Future Work

This pipeline will be finetuned and eventually used to classify all 1.6 million volumes added to the HTDL since the initial NovelTM dataset was generated, which will expand the pool of English-language fiction open to text and data mining, and document a reproducible process to continue to update the dataset as the HTDL grows. The train/test data and results of this project also hold possibility for further exploration of different classification approaches, such as ensemble models that include classifiers trained specifically on edge cases, such as memoir.

# Bibliography

**Bucher, Rolf.** *Classification of Fiction Genres : Text Classification of Fiction Texts from Project Gutenberg*, 2018. http://urn.kb.se/resolve?urn=urn:nbn:se:hb:diva-16007.

**Gupta, Rachna.** "Classifying Fiction and Non-Fiction Works Using Machine Learning." *Student Publications & Research*, October 29, 2019. https://digitalcommons.imsa.edu/student_pr/46.

**Jacob Jett, Boris Capitanu, Deren Kudeki, Timothy Cole, Yuerong Hu, Peter Organisciak, Ted Underwood, Eleanor Dickson Koehl, Ryan Dubnicek, J. Stephen Downie** (2020). *The HathiTrust Research Center Extracted Features Dataset (2.0)* . HathiTrust Research Center. https://doi.org/10.13012/R2TE-C227

**Miller, David P.** "Out from Under: Form/Genre Access in LCSH." *Cataloging & Classification Quarterly* 29, no. 1–2 (June 1, 2000): 169–88. https://doi.org/10.1300/J104v29n01_12.

**"NovelTM – .Txtlab @ Mcgill."** Accessed November 4, 2022. https://txtlab.org/category/textminingthenovel/.

**Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al.** "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12, no. 85 (2011): 2825–30.

**Short, Matthew.** "Text Mining and Subject Analysis for Fiction; or, Using Machine Learning and Information Extraction to Assign Subject Headings to Dime Novels." *Cataloging & Classification Quarterly* 57, no. 5 (July 4, 2019): 315–36. https://doi.org/10.1080/01639374.2019.1653413.

**Underwood, Ted, Patrick Kimutis, and Jessica Witte.** 2020. "NovelTM Datasets for English-Language Fiction, 1700-2009." *Journal of Cultural Analytics* 5 (2). https://doi.org/10.22148/001c.13147.