



Modular Bibliographical Profiling of Historic Book Reviews

RESEARCH PAPER

MATTHEW J. LAVIN 

ubiquity press

ABSTRACT

This paper examines different methods of predicting bibliographical details (e.g. author, title, and publisher) of books under review in a corpus of approximately 1,100 historical book reviews. The dataset is comprised of book reviews from ProQuest's American Periodicals Series (APS). This kind of bibliographical profiling is often characterized as a Natural Language Processing (NLP) or Named Entity Recognition (NER) task, but it can be more specifically described as a two-part Named Entity Linking (NEL) task, beginning with a feature extraction stage followed by one of several available matching or classification methods. An attempt has been made to formalize constraints for modular bibliographical profiling (MBP) and shed light on some important choices that are often glossed over or obscured by digital humanities practitioners. Applying these constraints, the paper evaluates combinations of feature selection (naive bag-of-words [BOW], rule-based feature extraction, and NER using a pre-trained model) with a standardized similarity-based matching strategy (cosine similarity). All tasks are performed on derived text data (term frequency tables), so that data can be shared and all methods can be used on materials available only in non-consumptive formats. These comparisons suggest that naive BOW can perform quite robustly, and that using even a basic pre-trained NER model in conjunction with a BOW approach may reduce false positives.

CORRESPONDING AUTHOR:

Matthew J. Lavin

Data Analytics, Denison
University, Granville, Ohio,
United States

lavinm@denison.edu

KEYWORDS:

book reviews; cultural
analytics; information
extraction; entity linking;
natural language processing;
named entity recognition

TO CITE THIS ARTICLE:

Lavin, M. J. (2024). Modular
Bibliographical Profiling of
Historic Book Reviews. *Journal
of Open Humanities Data*, 10:
26, pp. 1–14. DOI: [https://doi.
org/10.5334/johd.183](https://doi.org/10.5334/johd.183)

1 CONTEXT AND MOTIVATION

Recent work in cultural analytics (CA) has demonstrated the interpretive payoffs of analyzing historic book reviews at scale, as well as the desirability of shared datasets/corpora to support this inquiry (Boot, 2013; Hegel, 2018; Underwood & Sellers, 2016; Sharma et al., 2020; Sinykin, So, & Young, 2019; Walsh & Antoniak, 2021). A core aspect of this research is establishing a conceptual relationship between a review and the work or works that review is discussing. This Named Entity Linking (NEL) task is not often described. Historic book reviews are often to be found in large-scale digital collections of journals, magazines, and newspapers, which can make a subset of reviews feel already assembled and available for analysis. Platforms like the American Periodicals Series (APS) cannot represent important contexts of publication, circulation, and reception, as well as their own histories of assemblage and data curation. With all such collections, it is easier to analyze what is found in the collection than to account for what might be missing or excluded. These and other factors contribute to an environment where important considerations are given little or no attention in print, including but not limited to:

1. How reviews are categorized, and whether similar genres such as announcements, previews, longform criticism, etc. are excluded.
2. What criteria is used to determine if two reviews are responding to “the same book”? When do two versions of title become different enough that they should be distinguished from one another?¹
3. What criteria is used to determine authorship? When does editorial labor, translation, etc. rise to the level of authorship? Can the computational approach determine if two different names refer to one author, or if two identical references are different authors with similar names?
4. Can the computational approach distinguish between a work or author being reviewed and a work or author mentioned in a review?

Complexities like these can be tied directly to decisions at the level of code. Whether inheriting defaults or making overt decisions, scholars’ seemingly inconsequential actions could affect computational results and, by extension, their inferred conclusions. Matching author names on strings of text, for example, could amplify underlying gender bias in a corpus since women have been more likely than men to change their names upon marriage. More subtly, recognizing a corpus’ most famous authors slightly more consistently than less famous authors can also be source of distortion, since the most famous authors in many corpora are more likely to be male than female. When text processing algorithms are applied at increasing scales with less hand correction, these biases have the potential to be further amplified.

This article considers several approaches to the NEL task associated with book reviews, with a primary focus on modularizing the two core phases of the task: feature extraction and matching. Three Modular Bibliographical Profiling (MBP) experiments are attempted: (1) naive bag-of-words (BOW) matching; (2) rule-based feature extraction; and (3) feature extraction using a pre-trained Named Entity Recognition (NER) model. Performance benchmarks including recall, precision, and f1 scores are reported and discussed for all three approaches. The paper closes with recommendations for how these findings might inform large-scale computational analysis of book reviews.

2 LITERATURE REVIEW

2.1 BOOK REVIEWS AND NATURAL LANGUAGE PROCESSING

Within NLP, identifying details of the book or books focused on in a book review can be framed as a citation extraction or NER task, but both frames have important shortcomings or limitations. Lessons learned from both domains, meanwhile, are relevant to bibliographical profiling. Citation extraction, which focuses on automated recognition and classification of in-text citations and references, is arguably a less unified area of NLP research than NER, as there are stakeholders in computer science and various areas of application (Iqbal et al., 2021). However, the task of

1 Understanding bibliographical models like FRBR helps explain the complexity of this question (*Functional Requirements for Bibliographic Records: Final Report*, 2009).

profiling a reference using interdependent information—title, author name, source periodical—is similar to profiling a book review. NER focuses on identifying entities such as institutions, people, places, dates, dollar amounts, events, works of creative expression, etc. The body of methods literature on NER is large, and applications to digital humanities (DH) domains, especially by digital historians, are widespread (Ehrmann et al., 2024). Recent methods work in NER has focused on leveraging deep learning (Al-Moslmi et al., 2020; Li et al., 2020; Yadav & Bethard, 2019; Sevgili et al., 2022). Recent applications in digital history have focused on evaluating the suitability of novel machine learning (ML) architectures (e.g., BiLSTMs, transformers) to historical materials and using models with character and sub-word information to address “spelling variations and OCR errors” (Ehrmann et al., 2024, p. 31). Insights drawn from tackling these problems are likely to be relevant to profiling historical book reviews.

2.2 BOOK REVIEW ANALYSIS IN DIGITAL HUMANITIES

In DH contexts, applications of NER and citation extraction techniques appear to be widespread, and profiling book reviews appears to be a growing area of interest. There have been several DH articles in the past five to seven years that appear to depend on extracting information from book reviews or book reviews indices like the *Book Review Index* (Hegel, 2018; Sinykin et al., 2019; Underwood & Sellers, 2016). For the most part, in DH scholarship, NLP methods are described sparingly or not at all. Code and raw data are shared regularly by some scholars and inconsistently or not at all by others. In some cases, copyright or licensing restrictions make it difficult to share these materials. In other cases, procedural details are omitted or glossed over because they are framed as self-evident or unimportant. This is especially the case when DH scholarship is published with a “traditional” humanities audience in mind (e.g., in scholarly journals and scholarly monographs not intended specifically for DH or CA audiences). Some of these tasks might be computationally uncomplicated, but lack of detail in the literature generally makes it difficult to assess the complexity of both the tasks performed and solutions applied.

2.3 ANALOGOUS WORK IN DIGITAL HUMANITIES

How the “mechanical details” of computational inquiry relate to the theoretical or critical work of DH is a subject of much discussion (Drucker, 2017). Whereas many NLP and ML papers are foremost concerned with methodological novelty and performance benchmarks, DH scholarship must consider well-established and novel approaches, with a focus on adapting them to humanities contexts and interrogating the values they convey. Computational methods prompt practitioners to specify epistemic assumptions, often in the form of data curation and writing code. In various areas, DH scholarship has engaged closely with computational methods and considered consequences beyond those typically prioritized in computer science. Such scholarship has striven to understand the technical and theoretical components of methods well enough to identify cases where seemingly trivial decisions can have crucial downstream implications. As stated above, perhaps the closest analogous area is the application of NER in the context of digital history, where Ehrmann et al. advocate that “the next generation of historical NER systems” prioritize transferability “across historical settings”; “more systematic and focused” NER methods; using gold standards and of shared tasks for greater comparability; developing finer-grained historical NER; and more resource sharing (Ehrmann et al., 2024, p. 31). MBP is conceived as a scholarly intervention situated between theory and practice, and hopefully contributing to both.

3 DATASET DESCRIPTION

Object name Book review data for “Modular Bibliographical Profiling of Historic Book Reviews”²

Format names and versions txt, CSV

Creation dates 2019-08-01–2023-11-1

Dataset creators Eva Bacas, University of Pittsburgh (data curation, software), Zoe Pratt, Denison University (data curation, software); Thao Chu, Denison University (data curation, software)

² This dataset can be accessed and cited using the following DOI: <https://zenodo.org/doi/10.5281/zenodo.10092558>. Replication files (data and Python code) can be found at <https://github.com/mjlavin80/modular-bibliographical-profiling/>.

A sample of reviews from ProQuest's APS was used to evaluate all NEL tasks. The APS online is a robust source for this study for several reasons. Typically, newspaper and magazine digitization involves scanning page images of bound periodicals or reels of page images on microfilm. Optical Character Recognition (OCR) is performed either on a page-by-page basis, or after segmenting periodical content into articles. Digitized serials are often subject to copyright restrictions and, as a result, only non-consumptive text features such as document-term frequencies can be shared. When periodicals appear in collections like Hathi Trust, therefore, one would need to segment content into individual reviews and redo OCR to derive text data that can be used to conduct all MBP experiments. When content from periodicals is segmented, the digital content is often lacking metadata to reliably identify book reviews, determine whether a review focuses one or more than one book, or extract any information about the book or books being reviewed.

The APS was originally a microfilm collection created by University Microfilms International (UMI) in 1973 and expanded circa 1979. University Microfilms was founded in 1939 as a publisher of doctoral dissertations. In the 1980s, UMI began using the brand name ProQuest for use with its CD-ROM products, many of which collected and stored materials previously created for microfilm. In the late 1990s, UMI announced a new "online information service" called ProQuest Direct and, in 1998, it launched the Digital Vault Initiative ("Addenda", 1996). The APS collection was subsequently made available as the APS Online (Jacso, 1998). It can be assumed that, around this time, APS content was digitized, segmented into separate PDF files for each article, and described with additional metadata to facilitate search and browse functionality. The online database includes metadata pertaining to different search fields. Metadata values are often blank for a particular record and, in some cases, appear to have been inferred programmatically, as suggested by relatively high error rates (Common Field Codes, n.d.). Table 1 describes some of these fields:

VARIABLE	DATA TYPE	NOTES
author	categorical/string	no URIs or personography in source XML
title	categorical/string	
language	categorical/string	
subjects	controlled vocabulary	
source periodical	categorical/string	
document type	controlled vocabulary	e.g., advertisement, commentary, illustration, news, obituary, and review
source type	controlled vocabulary	e.g., newspaper, magazine, and trade journal
publication date	date string or integer	source XML both formats
abstract included	yes/no	designed with other Proquest platforms in mind?
peer reviewed	yes/no	designed with other Proquest platforms in mind?
scholarly	yes/no	designed with other Proquest platforms in mind?

Table 1 ProQuest metadata fields.

To help ensure a well-balanced sample, a random selection of articles tagged as reviews (1880–1925) was used. All selected articles were coded by hand to identify single-focus reviews.³ Table 2 summarizes the counts for each category in the initial sample of 2,155 items tagged as reviews.

³ Articles were excluded if they were categorized as non-book-review content, reviews focusing on more than one book, reviews shorter than 200 words, or reviews with a missing PDF file. Clusters of reviews digitized as one article (common with newspapers) were also omitted.

LABEL	COUNT
brief	82
cluster/multi	731
no_pdf	2
not_review	247
single_focus	1093
total	2155

Reviews established as ‘single-focus’ were subsequently coded with labels for author, title, and publisher, as well as minimal genre tags to separate reviews of drama, fiction, non-fiction, and poetry from one another. [Table 3](#) summarizes genre counts.

LABEL	COUNT
nonfiction	801
fiction	226
poetry	34
drama	18

Table 2 Hand-labeled categories for sample items tagged as reviews.

Table 3 Hand-Labeled genres for sample items tagged as single-focus reviews.

4 METHOD

4.1 EXPLORATORY DATA ANALYSIS

The methods explored in this paper differentiate between (1) approaches based on any Boolean (True/False) string matching criteria (literal, fuzzy, or pattern matching) and (2) approaches where a feature extraction phase is paired with a metric-based similarity scoring phase (MBP). As Warren et al. have argued, confidence estimates are often preferable when dealing with records of unequal informational value and are “better suited to the grey areas of humanistic research” (Warren et al., 2016). Using a string-matching approach, one might loop through a list of authors, titles, or publishers and classify any in-text match of a string as a mention of the corresponding entity. The dataset used in this article generates a list of 1,155 unique author surnames for 1,093 reviews. Of these surnames, approximately 90% can be found in their corresponding reviews, but that leaves about 10% of surnames unmentioned in the review text. With book titles, one can tokenize the text, control for capitalization, and limit each title to a four-gram with stopwords included or excluded and find approximately 70% of book titles in their corresponding reviews (out of 1,087 distinct title strings). With publisher names, repeats are much more common even in a small sample, so the dataset yields 258 distinct strings or n-grams (after controlling for non-standard tail words like “corp” and “inc”). NLP-based matching can discover more than half of these publisher names in their corresponding review text. In some cases, authors, titles, or publishers are simply not mentioned in book reviews. In other cases, OCR errors lead to missed matches or false negatives. [Table 4](#) summarizes match rates as well as false positive counts.

ENTITY	PERCENT MATCHED	RECALL	PRECISION	F1 SCORE
Author surnames	90.11	0.54	0.16	0.13
Title n-grams (4 max)	69.45	0.41	0.40	0.20
Title n-grams (4 max, no stopwords)	68.16	0.41	0.17	0.12
Publisher n-grams (4 max)	51.62	0.68	0.47	0.28
Publisher n-grams (4 max, no stopwords)	52.91	0.69	0.46	0.28
Publisher n-grams (4 max, no stops or tails)	57.63	0.71	0.39	0.25

Table 4 Performance measures for authors, titles, and publishers.

Among author, publisher, and title, author surnames are the most likely to be matched in reviews, at a rate of approximately 90%.⁴ If we look closer at recall, precision, and F1 scores, there are several surprises. First, the number of false positives drives down all three performance statistics. False positives occur most often when the entity being matched overlaps with commonly used words or phrases, as with author surnames like Long, Day, Church, and London; book titles like *Poems*, *Missouri*, and *Bliss*; or title n-grams like “history of the [noun]” or “the story of a [noun]”. They can also occur in the case of entity overlap or ambiguity, such as when a book titled *Balsac* (a biography of the author) boosts the likelihood of Balsac being the author under review or a reference to *Appleton’s Encyclopedia* boosting the likelihood of Appleton being the book’s publisher (Kemp, 1909).

Hopefully it is clear that simplistic NLP-based matching approaches make certain kinds of errors more likely than they should be. Perhaps most importantly, the ceiling for false positives is raised. With traditional classification or matching tasks, false positives cannot be greater than the number of observations (N). With simplistic string matching, each review can produce one correct title or publisher, one or more correct authors, and as many false positives as the number of candidates one uses for matching. Allowing more fuzziness to regular expressions or N-gram comparisons may seem like an appealing way to reduce false negatives, but one should expect false positives to increase at a greater rate than true positives. As fuzziness is widened, true positives tend to follow a pattern of diminishing returns, while false positives will either stay steady or increase, with an upward bound of the number of tokens or n-grams in a document.

4.2 MODULAR BIBLIOGRAPHICAL PROFILING

This study is designed to control for known sources of uncertainty and make gains in book review profiling related to feature extraction and matching book review text to author, title, publisher triads. The most common errors associated with reconciling book review features to bibliographical records or identifiers seem to be best isolated by eliminating any errors related to incorrect page segmentation and misidentifying non-review content as book reviews. Focusing on reviews with only one correct bibliographical target simplifies performance evaluation while maintaining opportunities for generalization. The tasks of page segmentation, review identification, and review type classification are crucial to analysis of historical book reviews, but they warrant full length studies of their own.

In this study, the following procedure is employed:

1. All single-focus book reviews in the dataset (N = 1,093) are used for each task
2. Each review’s full text is pre-processed (e.g., tokenization, punctuation preserved or removed, capitalization preserved or ignored)
3. A feature extraction or selection strategy is employed to isolate text features
4. A list of bibliographical entity candidates is derived by taking the correct labels for all authors, titles, and publishers from the metadata and reducing each triad to a single entry of raw text
5. Reviews and triads alike are converted to term-frequency tables
6. A similarity measure is employed to compare each review to each triad
7. For each review, a ranked list of most similar triads is returned
8. If the top-ranked triad is the correct entry, the match is considered correct
9. If the top-ranked triad is not the correct entry, a false positive recorded for the matching record and a false negative is recorded for the true match
10. If the match is incorrect, the rank of the correct match is used for “in the ballpark” statistics (below)

No partition of training and test set is used here because every bibliographical record in the dataset is unique. Instead, all true labels are in the candidate set, and each matching task

⁴ With some reviews containing more than one author, it is also important to count to the number of reviews that matched 100% of their authors correctly, and this statistic is roughly on par with total author names matched at 89.8%.

has a number of potential false positives equal to the total number of candidate labels, minus one (1,092). When conducting this task using random guessing, we would expect an accuracy rate of about 0.001%. Table 5 details the feature selection approaches that are considered.

NUMBER	FEATURE STRATEGY
1	Naive BOW
2	Rule-based entity extraction
3	Pre-trained NER

Table 5 Feature selection strategies under evaluation.

The feature extraction phase is designed to be modular, so that any feature extraction can be swapped for another without changing any other aspects of the code.

Strategy 1 (naive BOW) involves tokenizing all book reviews and all author-title-publisher triads, and then transforming them into a single BOW model such that the term-frequency matrix includes every unique token from all reviews and all triads.

Strategy 2 (rule-based entity extraction) uses a complex set of rules to find likely author surnames, and then infer forenames based upon those surnames. All possible titles and publishers were reduced to n-grams of non-stopword tokens with a maximum length of four (e.g., “Gone with the Wind” produces the tokens “gone” and “wind” and “From the Mixed-Up Files of Mrs. Basil E. Frankweiler”, produces the tokens “mixed-up”, “files”, “mrs”, “basil”).⁵

Strategy 3 (pre-trained NER) began by running NER on all reviews using a pre-trained model (Spacy’s “en_core_web_trf”). Entities belonging to the categories of Person, Geopolitical Entity, Nationality, Organization, Facility, Event, Location, Product, Work of Art, and Law were retained and all other recognized entities were dropped. As with Strategies 1 and 2, a BOW model was generated with a vector space that contained all tokens from the list of qualifying recognized entities.

For the matching portion of the experiments, the same process was used for all three strategies. From within the derived vector space, every book review vector was compared to every author-title-publisher triad, and a ranked list of review-to-triad cosine similarities was generated. If the correct record ranked first in the list of similarity scores, the match was scored as correct.

5 RESULTS AND DISCUSSION

As Table 6 shows, accuracy scores for this task were generally high, despite the encumbrance of using only non-consumptive models. The parameters of the study, by design, restrict the number of false positives and false negatives such that the recall score provides a strong initial assessment of each approach’s overall performance. To augment these scores, precision and f1 score are also provided but there are 1,093 distinct class labels (each review is its own class), so only pooled recall, precision, and f1 scores are provided. To derive these scores, metrics are calculated for each label and then averaged, with weighting for the number of true instances for each label in case of any class imbalances.

NUMBER	FEATURE STRATEGY	RECALL	PRECISION	F1 SCORE
1	Naive BOW	86.19	81.69	83.01
2	Rule-based entity extraction	83.97	78.64	80.21
3	Pre-trained NER	84.71	79.25	80.92

Table 6 Performance measures for bibliographical records.

Examining the performance statistics from Table 6, the biggest surprise is how well naive BOW performed, with the highest recall, precision, and f1 scores. The pre-trained NER strategy performed second best of the three, and rule-based extraction was the worst, but only by a slim margin. All three models could be strong candidates for a ball-parking task but, as it

⁵ See <https://github.com/mjlavin80/modular-bibliographical-profiling/blob/main/Classification-Experiments-BOW.ipynb>.

stands, naive BOW would have to be favored because it performs best and requires the least computational pre-processing.

To examine performance a bit more closely, Figure 1 provides an overview of how many reviews out of 1,093 could be considered near matches. Since there is no objective threshold that one might call “almost correct”, the plot begins with a relatively small threshold of nearness (if the true label is within the top five matches for the review) and shows the effect of increasing this threshold in steps of five (top five, top ten, etc.) from five to fifty.

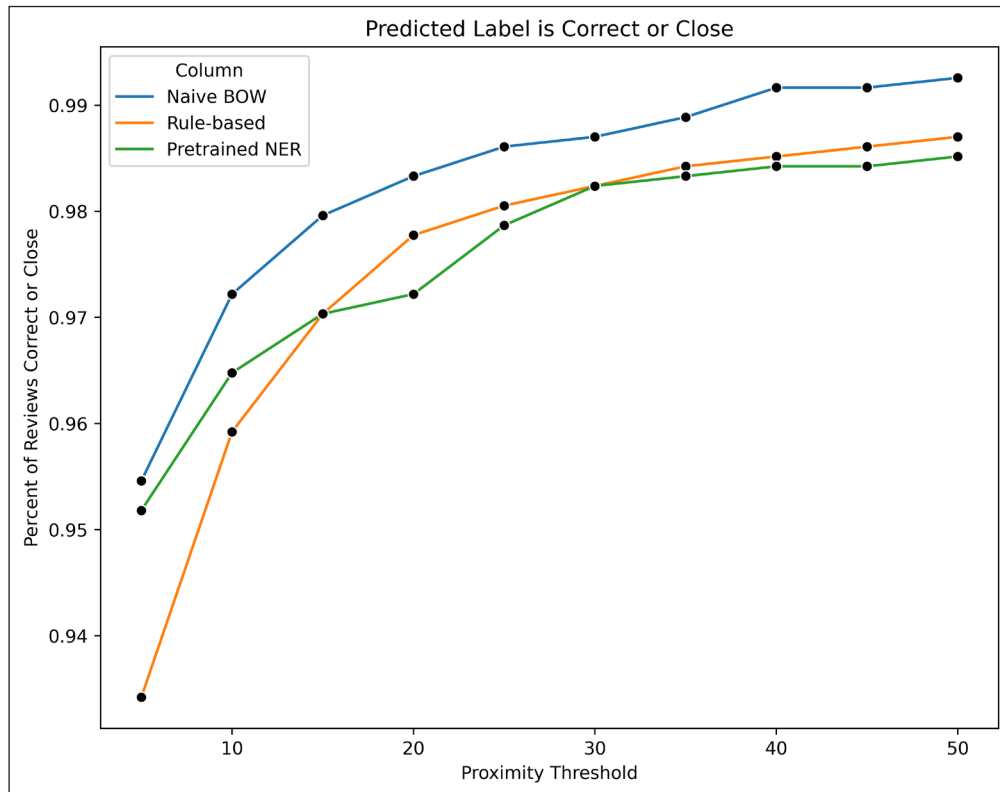


Figure 1 Incrementing threshold effect on match or near match rate.

Overall, this figure shows if any of the feature selection strategies performs better as a ball-parking strategy than an exact match strategy. As Figure 1 shows, however, naive BOW has the highest rates at all threshold levels; that is, it matches the most bibliographical records correctly (86.19%), it has the greatest number of bibliographical records within the top five matches (93.42%), the top ten matches (95.92%), etc. Rule-based entity extraction and the pre-trained NER strategies are more muddled, with the lines crossing one another as the threshold increases. Information of this kind would be especially useful if one were hoping to use computational methods to select likely candidates for the Bibliographical Profiling task and then use hand correction to ensure maximum accuracy. In the context of a Human-in-the-Loop (HITL) platform, being able to select the correct entity from a list of finalists would drastically speed up the encoding process, even if the correct entity were absent from that list occasionally.

The structure of the matching task limits each row to one correct or incorrect answer, but there is nothing to prevent any one bibliographical entity from being the selected false positive over and over again, which means that false positives can be analyzed collectively. Reviews that tend to pop over and again up as false positives may even have attributes that might make their status as “false positive magnets” more likely. First, it can be determined whether the three feature extraction strategies produce similar patterns of false positives. As Table 7 shows, the naive BOW and rule-based strategies tended to similar patterns of false positives, whereas the pre-trained NER strategy had a weak, negative correlation with both naive BOW and rule-based strategies.

	NAIVE BOW	PRE-TRAINED NER	RULE-BASED
Naive BOW	1	-0.17	0.82
Pre-trained NER	-0.17	1	-0.15
Rule-based	0.82	-0.15	1

Table 7 Pearson correlation matrix for false positive rates by feature strategy.

Since their performance rates are so close, it is surprising that the pre-trained NER strategy produces such different patterns of false positives. As Table 8 shows, false positive magnets using the pre-trained NER strategy appear to be relatively sparse records with the generic nouns or place names in the title.

AUTHOR	TITLE	PUBLISHER	DATE	PRE-TRAINED NER	NAIVE BOW	RULE-BASED
Abraham Cahan	<i>Yekl: A Tale of the New York Ghetto</i>	Appleton	1896	11	1	1
John D. Long	<i>The New American Navy</i>	Outlook	1903	4	1	3
James Smetham	<i>Letters of James Smetham</i>	Macmillan	1902	4	0	0
Mrs. Schuyler Van Rensselaer	<i>One Man Who Was Content</i>	Century	1897	0	12	16
Juliet Adams	<i>My Literary Life</i>	Appleton	1904	0	5	5
E.F. Benson	<i>Mrs. Ames</i>	not listed	1912	0	9	0
Cuthbert Wright	<i>One Way of Love</i>	Elkin Mathews	1915	0	4	5

Table 8 Summary of largest false positive magnets.

Meanwhile, naive BOW and rule-based strategies produce more false positive magnets overall, and they seem to occur when author-title-publisher triads have relatively few tokens, especially when book titles are short or filled with vague, high-frequency terms. More likely than not, false positive magnets under these conditions are not so much providing a strong match with many book reviews but are rather providing a mediocre match that ranks highly in the absence of an obvious prediction.

This line of analysis raises the question of whether a “best match” similarity score itself can predict a false positive. That is, if a best match is weak, is it more likely to be a false match? This would be a very simple and straightforward basis from which to flag ambiguous matches and would be easy to refine by adding additional decision criteria. Evaluating such an approach begins with comparing measures of central tendency and the distributions of top-ranked match scores. For all three strategies, as Table 9 shows, true positives have higher means, higher minimum values, higher maximum values, and higher cutoffs in each quartile than the set of false positives derived from the same strategy.

	NAIVE BOW		PRE-TRAINED NER		RULE-BASED	
	TRUE POS	FALSE POS	TRUE POS	FALSE POS	TRUE POS	FALSE POS
mean	0.28	0.20	0.48	0.32	0.37	0.25
min	0.09	0.06	0.08	0.08	0.14	0.07
25%	0.22	0.15	0.37	0.23	0.30	0.20
50%	0.27	0.18	0.47	0.31	0.37	0.23
75%	0.33	0.23	0.58	0.40	0.44	0.29
max	0.60	0.52	0.98	0.75	0.71	0.66

Table 9 Measures of central tendency for top-ranking similarity scores.

The absolute values for these measures, however, appear quite variable. Going further, the question of how well the continuous similarity scores can be separated into the binary categories of “likely true positive” and “likely false positive” can be assessed by using a logistic regression model as a diagnostic tool. When a logistic regression model is trained using top ranking cosine similarity scores as the independent variable and the binary labels of “true positive” and “false positive” as the dependent variable, a baseline accuracy of 80–85% (86% for naive BOW, 80% for pre-trained NER, and 84% for Rule-based Extraction) is achieved. To ensure that these scores have diagnostic value, a constraint is added requiring the model to assign the label of “false positive” to about 20% of the records; otherwise, labeling all records as “true positives” would yield an accuracy rate equal to the underlying strategy’s accuracy

rate. Instead, the goal is a model that has a chance of high overall accuracy and balanced precision and recall scores for both labels.

More significant than overall accuracy, uncertainty in all three models is predominantly found in the lowest fifth of the data. That is, if the lowest 20% of the similarity scores are grouped, this cluster is composed of approximately half true positives and half false positives. The upper 80% of the similarity scores tend to be more than 90% true positives.

As Table 10 demonstrates, the strategy of labeling all cosine similarity scores in the bottom 20% of the sample “needs audit” increases the precision of the remaining 80% of the data. If this approach is used to establish a cosine similarity score threshold and consider all predictions below this level to be false, precision can be maximized with a relatively small effect on recall (that is, allow more false negatives in order to minimize false positives). As Figures 2, 3, and 4 show, this approach could capture approximately 80% of all correctly matched reviews while maintaining precision rates above 90%.

STRATEGY	QUARTILE(S)	COSINE SIMILARITY RANGE	ACCURACY
Naive BOW	Lower	0.06–0.19	60.19
	Upper Three	0.19–0.60	92.82
Rule-based	Lower	0.00–0.26	51.39
	Upper Three	0.26–0.71	91.55
Pre-trained NER	Lower	0.00–0.32	43.06
	Upper Three	0.32–0.98	89.12

Table 10 Uncertainty by strategy, lower quartile vs. upper three quartiles.

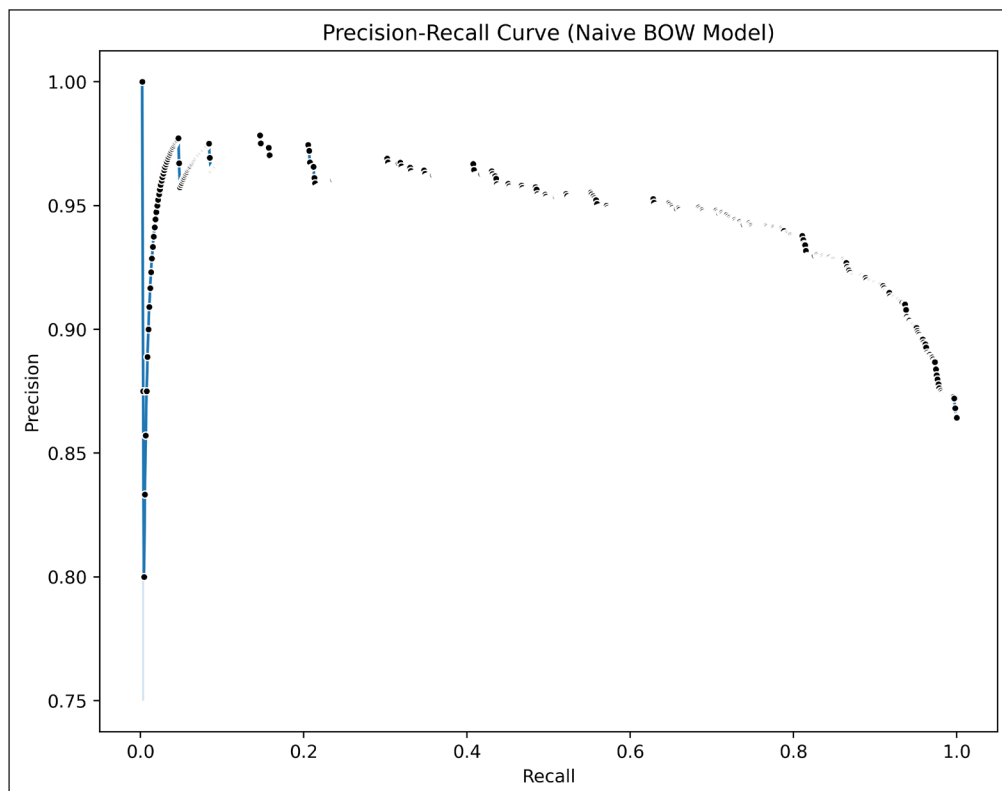


Figure 2 Precision-recall curve (naive BOW model).

This analysis suggests that even a simple approach to predicting weak matches could be quite effective. There is not enough data in the current sample to use a cross-validation strategy, so performance may be inflated, but the biggest remaining question is whether strong and weak prediction cosine similarity scores will distribute similarly when there are many more reviews under consideration and many more author-title-publisher candidates in the mix. Due to limitations of the current study, this question must remain unanswered, but it is a logical next step.

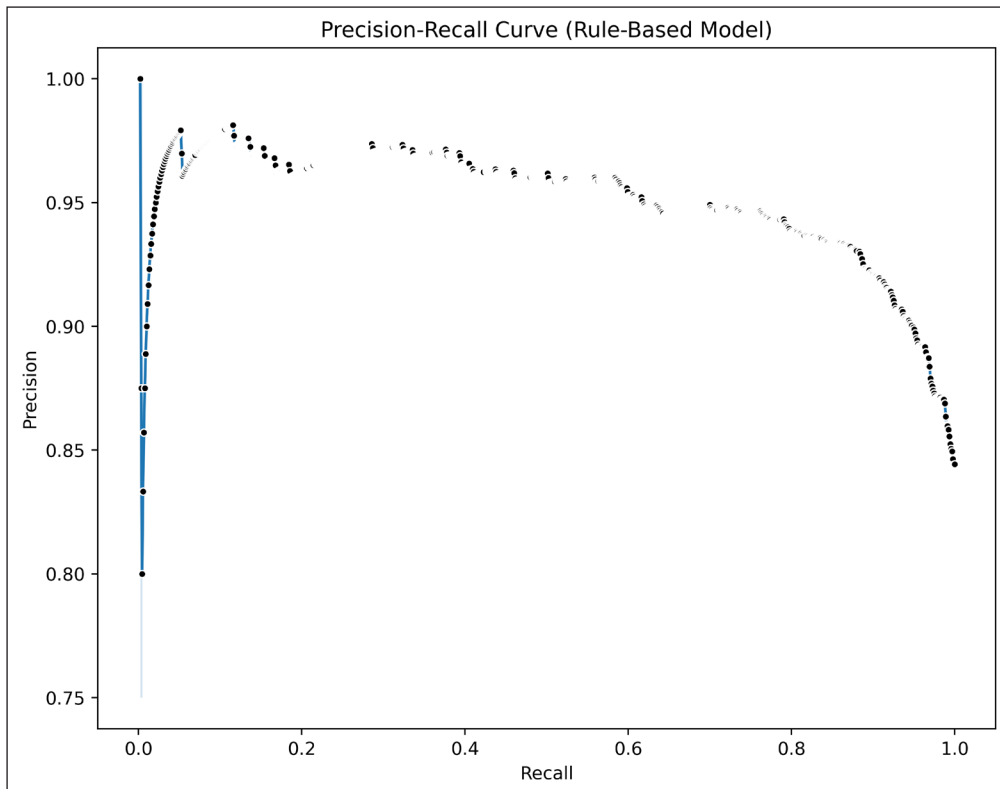


Figure 3 Precision-recall curve (rule-based model).

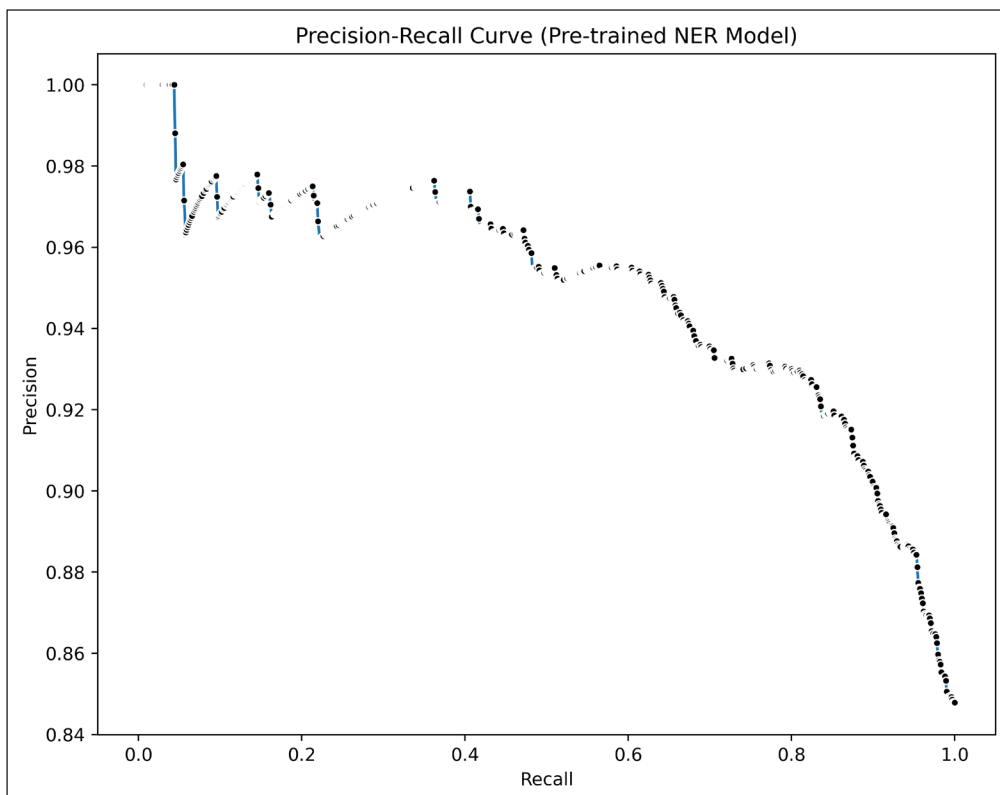


Figure 4 Precision-recall curve (pre-trained NER model).

6 IMPLICATIONS AND APPLICATIONS

6.1 DATASET LIMITATIONS AND REUSE POTENTIAL

The dataset used for this paper is put forth as a well-balanced (albeit relatively small) sample of single-focus reviews from Proquest's APS database originally published in prominent weekly and monthly publications in the United States between 1880 and 1925. The constraints of this sample imply several limits to the broader representativeness of the data, especially temporal, cultural, and linguistic representativeness. All reviews are written in English, and it can be expected that at least some of the performance statistics reported here would differ if similar methods were applied to reviews in other languages.

This dataset should prove to be useful in a range of classroom contexts (review analysis, testing NLP methods on shorter texts, etc.) and can be used to benchmark the performance of additional approaches to entity profiling, with the caveat that, with each reuse, the risk of overfitting increases, which would potentially prevent refinements from generalizing to other reviews. Finally, this dataset may aid in future efforts to locate reviews in digitized periodicals and extract bibliographical entities from them, which could result in more representative datasets over time.

6.2 MODULARITY AS A TEMPLATE

The dataset used for this study is structured in order to function as one component of a larger pipeline. This workflow is intended to act as a template for future analysis of this sort, as well as work on this dataset and the creation of comparable datasets. A modular approach such as the one laid out above should be given serious consideration in the context of large-scale bibliographical profiling of book reviews, and beyond. This kind of approach has both procedural advantages and a strong potential to improve performance (measured in terms of recall and precision). These benefits include:

1. **Modularity:** The approach is designed to create discrete partitions between the feature extraction phase and the similarity scoring phase so that any feature extraction strategy (including strategies not explored in this paper) could be used interchangeably, as could additional similarity measures.
2. **Simplicity:** Increasingly intricate rules are eschewed in favor of more general NLP strategies. The dataset eliminates several well-known confounds for book review analysis and provides a basis to compare the performance of the feature extraction and entity linking steps of the process.
3. **Extensibility:** Items 1 and 2 should make these profiling approaches easier to adapt and extend. Additional refinements are expected and encouraged, with the caveats described in section 3.1.
4. **Non-Consumptive Formats:** All of the above methods can be used on corpora of book reviews where only non-consumptive format is available. Periodicals will still need to be segmented and OCR'd at the article level, but it is hoped that this work will motivate entities like ProQuest, JSTOR, and HTRC to segment more periodicals and/or release more non-consumptive formats representing the article level of periodicals.
5. **Books as Bibliographical Entities:** The method established above cannot create bibliographical records that do not exist by “mixing and matching” author, title, and publisher combinations. Adding this control creates the opportunity to establish a performance baseline that can be manipulated in future analysis.
6. **Novelty Detection:** Building on item 4, a subsequent study could arbitrarily remove an incrementing percent of true labels from the candidate set (or introduce an incrementing percent of “nuisance candidates”) to assess how predictably performance decreases under new uncertainty. It could also help researchers identify obscure or as-yet-unknown authors, publishers, or books. This potential is especially important when one considers the fact that matching bibliographical records from predetermined could be a source of bias amplification, specifically in the form of confirmation bias.
7. **Applications Beyond Book Reviews:** It is hoped that this work will help motivate and enable research occupying a middle space between work traditionally categorized as theory/methods or practice/application. Opportunities to extend the “praxis work” of DH are abundant.

ACKNOWLEDGEMENTS

I wish to thank Eva Bacas (University of Pittsburgh), Zoe Pratt (Denison University), and Thao Chu (Denison University) for their participation in this project. They worked as paid undergraduate research assistants and received their home university’s predetermined rate of pay for all undergraduate employees. All three students assisted with review labeling, data curation, and writing code. I would like to thank my wife Rebecca Lee for her support on this project and assistance with the manuscript. I am also indebted to ProQuest for providing access to OCR data, metadata, and raw pdf files comprising the APS database.

FUNDING INFORMATION

The University of Pittsburgh and Denison University played a part in funding this research in the form of paid undergraduate research assistants. University of Pittsburgh research funds were used to license APS data.


COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR INFORMATION

Matthew J. Lavin (conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing–original draft, writing–review & editing), Eva Bacas (data curation, software), Zoe Pratt (data curation, software); Thao Chu (data curation, software).

AUTHOR AFFILIATIONS

Matthew J. Lavin  orcid.org/0000-0003-3867-9138
Data Analytics, Denison University, Granville, Ohio, United States

REFERENCES

- Addenda.** (1996, June). *The New York Times*, D9. Retrieved from <https://timesmachine.nytimes.com/timesmachine/1996/06/03/issue.html>
- Al-Moslmi, T., Ocaña, M. G., Opdahl, A. L., & Veres, C.** (2020). Named entity extraction for knowledge graphs: A literature overview. *IEEE Access*, 8, 32862–32881. Retrieved 2023-11-02, from <https://ieeexplore.ieee.org/abstract/document/8999622/> (Publisher: IEEE). DOI: <https://doi.org/10.1109/ACCESS.2020.2973928>
- Boot, P.** (2013). The desirability of a corpus of online book responses. In *Proceedings of the Workshop on Computational Linguistics for Literature*. ACL.
- Common Field Codes.* (n.d.). Retrieved 2023-11-02, from <https://www.proquest.com/help/academic/ViewFieldCodes.html>
- Drucker, J.** (2017, May). Why Distant Reading Isn't. *PMLA*, 132(3), 628–635. Retrieved 2024-02-01, from <https://www.cambridge.org/core/journals/pmla/article/abs/why-distant-reading-isnt/757C1225CFDCF629FC2895C76DD747B0> (Publisher: Cambridge University Press) DOI: <https://doi.org/10.1632/pmla.2017.132.3.628>
- Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A.** (2024, February). Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys*, 56(2), 1–47. Retrieved 2024-02-01. DOI: <https://doi.org/10.1145/3604931>
- Functional Requirements for Bibliographic Records: Final Report.* (Tech. Rep.). (2009, February). Retrieved from <https://repository.ifla.org/handle/123456789/811>
- Hegel, A.** (2018). *Social Reading in the Digital Age* (PhD Thesis). UCLA.
- Iqbal, S., Hassan, S.-U., Aljohani, N. R., Alelyani, S., Nawaz, R., & Bornmann, L.** (2021, August). A decade of in-text citation analysis based on natural language processing and machine learning techniques: an overview of empirical studies. *Scientometrics*, 126(8), 6551–6599. Retrieved 2023-11-02. DOI: <https://doi.org/10.1007/s11192-021-04055-1>
- Jacso, P.** (1998, September). UMI's Digital Vault Initiative project. *Information Today*, 15(8), 15–16.
- Kemp, R. W.** (1909, April). The Letters of Mrs. James G. Blaine. *The Bookman; a Review of Books and Life (1895–1933)*, 29(2), 193. Retrieved from <http://search.proquest.com/docview/124748263/>
- Li, J., Sun, A., Han, J., & Li, C.** (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1), 50–70. Retrieved 2023-11-02, from <https://ieeexplore.ieee.org/abstract/document/9039685/> (Publisher: IEEE). DOI: <https://doi.org/10.1109/TKDE.2020.2981314>
- Sevgili, O., Shelmanov, A., Arkhipov, M., Panchenko, A., & Biemann, C.** (2022, January). Neural entity linking: A-survey of models based on deep learning. *Semantic Web*, 13(3), 527–570. Retrieved 2023-09-07, from <https://content.iospress.com/articles/semantic-web/sw222986> (Publisher: IOS Press) DOI: <https://doi.org/10.3233/SW-222986>
- Sharma, A., Hu, Y., Wu, P., Shang, W., Singhal, S., & Underwood, T.** (2020). The Rise and Fall of Genre Differentiation in English-Language Fiction. *CEUR Workshop Proceedings*, 97–114. Retrieved from <http://ceur-ws.org/Vol-2723/long27.pdf>

- Sinykin, D., So, R. J., & Young, J.** (2019). Economics, Race, and the Postwar US Novel: A Quantitative Literary History. *American Literary History*, 31(4), 775–804. (Publisher: Oxford University Press). DOI: <https://doi.org/10.1093/alh/ajz042>
- Underwood, T., & Sellers, J.** (2016). The Longue Durée of literary prestige. *Modern Language Quarterly*, 77(3), 321–344. (Publisher: Duke University Press). DOI: <https://doi.org/10.1215/00267929-3570634>
- Walsh, M., & Antoniak, M.** (2021). The Goodreads “Classics”: A Computational Study of Readers, Amazon, and Crowdsourced Amateur Criticism. *Journal of Cultural Analytics*, 4, 243–260. (Publisher: Department of Languages, Literatures, and Cultures). DOI: <https://doi.org/10.22148/001c.22221>
- Warren, C. N., Shore, D., Otis, J., Wang, L., Finegold, M., & Shalizi, C.** (2016, July). Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks. *Digital Humanities Quarterly*, 10(3).
- Yadav, V., & Bethard, S.** (2019, October). *A Survey on Recent Advances in Named Entity Recognition from Deep Learning models*. arXiv. Retrieved 2023-11-02, from <http://arxiv.org/abs/1910.11470> (arXiv:1910.11470 [cs]).

TO CITE THIS ARTICLE:

Lavin, M. J. (2024). Modular Bibliographical Profiling of Historic Book Reviews. *Journal of Open Humanities Data*, 10: 26, pp. 1–14. DOI: <https://doi.org/10.5334/johd.183>

Submitted: 13 November 2023

Accepted: 08 February 2024

Published: 18 March 2024

COPYRIGHT:

© 2024 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

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