Estimating the Date of First Publication in a Large-Scale Digital Library

David Bamman  
School of Information  
University of California, Berkeley  
dbamman@berkeley.edu

Michelle Carney  
School of Information  
University of California, Berkeley  
michelle.carney@berkeley.edu

Jon Gillick  
School of Information  
University of California, Berkeley  
jongillick@berkeley.edu

Cody Hennesy  
Doe Library  
University of California, Berkeley  
chennesy@library.berkeley.edu

Vijitha Sridhar  
Computer Science Division  
University of California, Berkeley  
vvsridhar@berkeley.edu

ABSTRACT

One prerequisite for cultural analysis in large-scale digital libraries is an accurate estimate of the date of composition of the text—as distinct from the date of publication of an edition—for the works they contain. In this work, we present a manually annotated dataset of first dates of publication of three samples of books from the HathiTrust Digital Library (uniform random, uniform fiction, and stratified by decade), and empirically evaluate the disparity between these gold standard labels and several approximations used in practice (using the date of publication as provided in metadata, several deduplication methods, and automatically predicting the date of composition from the text of the book). We find that a simple heuristic of metadata-based deduplication works best in practice, and text-based composition dating is accurate enough to inform the analysis of “apparent time.”

CCS CONCEPTS

• Information systems → Digital libraries and archives;

KEYWORDS

Digital libraries; bibliographic metadata; publication date prediction

ACM Reference format:
DOI: 10.475/123_4

1 INTRODUCTION

The rise of large-scale digital libraries—such as those by Google Books, the Internet Archive, and the HathiTrust—has enabled a range of work in cultural analytics over the past decade, helping provide the raw material for the historical analysis of genre [47], character [1], emotion [18], loudness [25], geographic attention [48] and much more.

For all of this work, it is important to have a rich understanding of a corpus prior to drawing conclusions about it; one important feature for understanding texts is their date of publication, since the primary variable in cultural analysis is often some quantity (such as word frequency) anchored specifically in time. For example, Michel et al. (2010) [32] was one of the first studies to make use of these extensive resources, and measure “fame” (among other quantities) by tracing the frequency of mention of a person’s name over the scope of their collection. Time is critically important for cultural analysis within these datasets, since arguments often hinge on exactly when a word was written, and criticisms may arise at the uncertainty of that information [37, 40].

For books, however, time can be measured in different ways. As the Functional Requirements for Bibliographic Records (FRBR) [20] model articulates, books can be viewed in several abstract categories, and each of those categories may have different temporal information associated with it—including the publication date of a specific edition (a manifestation, in FRBR terminology), and the first appearance of any expression of the work overall. The public domain texts of the HathiTrust, for example, include an edition of Jane Austen’s Pride and Prejudice published in 1920 by D. M. Dent & Sons as a reprint of an edition originally published in 1906; all of these texts are different FRBR expressions of Austen’s work, first published a century earlier in 1813. If we are using Austen’s text as a source for quantities in cultural history (such as the first appearance of a particular word), then we would likely associate the original date of publication of 1813 with the text, rather than the date of printing of any later editions. If, on the contrary, we are using Austen’s text to investigate questions of typography, then we may be more likely to associate the publication date of the edition. Different research questions require different associations of date.

This disparity between the observed date of publication for a specific printing and the (often unobserved) date of first publication is exacerbated for popular texts that undergo several reprints. Figure 1 illustrates this for the works of Jane Austen in the public domain of the HathiTrust; nearly all of Austen’s works in this digital library are published after her death in 1817, most nearly a century later.
we create three gold standard datasets of books drawn from the To help drive an empirical evaluation of the accuracy of several work methods for identifying the date of publication of a FRBR ods. Our contributions are the following: a/t_tributing two unlike books to be the same). failing to recognize duplicates, or errors of commission in falsely introducing errors of its own (either errors of omission in failing to recognize duplicates, or errors of commission in falsely attributing two unlike books to be the same).

In this work, we present an empirical evaluation of these methods. Our contributions are the following:

- We present a manually annotated, gold standard dataset for the dates of first publication for 2,706 books in the public domain of the HathiTrust, which we are publicly releasing under a Creative Commons license for others to use as a benchmark.
- We find approximately 10% of books of the HathiTrust have a difference of at least 10 years between their first publication date and observed publication dates in the metadata, suggesting the overall difference is large enough to warrant consideration.
- We find that when digital library metadata is comprehensive enough, simple deduplication-based methods are often accurate enough to substantially reduce this error.
- In addition to supplementing metadata-based assignment methods, estimating the date of publication from the content of a book can also yield insight into the determinants of temporal signatures in a text.

2 MANUAL DATING
To help drive an empirical evaluation of the accuracy of several methods for identifying the date of publication of a FRBR work, we create three gold standard datasets of books drawn from the HathiTrust paired with their manually identified date of first publication. From all books in the public domain of the HathiTrust, we draw three separate samples of approximately 1,100 books each.

- **UNIFORM.** We draw a uniform random sample of books from the public domain of the HathiTrust.
- **FICTION.** We draw a uniform random sample of books from the 102,349 works automatically identified to be fiction by Underwood 2014 [46].
- **STRATIFIED.** We draw a random sample of books from the public domain of the HathiTrust, but stratify the sample by observed publication date, so that we have a roughly equal number of books from each decade between 1750-1922.

We then attempt to manually identify the first publication date for each of these books using the process described below.

2.1 Process
Despite significant advances in implementing the FRBR model in bibliographic catalogs such as WorldCat [19], there is still no comprehensive, authoritative database that will merge or visualize a FRBR work with all of its known expressions and manifestations. To complicate matters, the bibliographic records that we depend on for documentation of publication dates were created according to standards and practices that have shifted over time [7], and have varied by institution [36].

In consideration of the above challenges, we consulted bibliographic records from a variety of sources, as well as digital books from a few major repositories, and assigned earlier dates of publication whenever there was reasonable documentary evidence to suggest that the year provided for a given book in our sample was not the earliest known edition. Any assigned date in our dataset should not be cited as an authoritative judgement on that book’s earliest publication, but rather as an indication that the bibliographic record suggests there was an earlier publication date than the edition at hand. The sources we consulted in the process were:

- Bibliographic records from the Nineteenth-Century Short Title Catalogue (NSTC), available in C19: The Nineteenth Century Index (ProQuest).
- Library catalog records from WorldCat (OCLC).
- Date and edition notes in the front matter and prefatory remarks from other digital editions of the title available in HathiTrust, Google Books, and Eighteenth Century Collections Online (Gale).
- Less frequently, and for more difficult titles, we referred to encyclopedias, descriptive bibliographies of specific authors, and/or book reviews from periodicals indexed in C19.

The NSTC unites records for English language printed works published between 1801 and 1918 from eight major libraries where, as noted in the introduction to the first series, cataloging standards varied widely [36]. While enumerative bibliographies such as the NSTC provide an imperfect or incomplete record of a book’s publication history [10, 11], book historians note their importance and relative accuracy for large-scale bibliographic scholarship [12, 45]. The majority of the bibliographic records in the NSTC were created after “distinguishing different editions of the same work had become the norm in bibliographic control” and before a shift in cataloging standards towards FRBR and a renewed focus on the work [7]. Most publication dates in the NSTC were accordingly

![Figure 1: Distribution of publication dates for books in the HathiTrust for whom Jane Austen is the author. Austen’s death is marked at 1817.](image-url)
drawn directly from a book’s title pages. Records for some undated imprints, however, appear to have used the nearest approximate beginning or middle year of a decade (e.g., 1850 or 1855) as an approximate publication date without documenting that fact [12].

Bibliographic records in WorldCat offer two distinct advantages to those in the NSTC: first, modern cataloging guidelines state to clearly document when dates used were approximate, based on any sources outside of the book itself, or represented dates other than that of the imprint (e.g., copyright) [15]. Second, WorldCat contains over 220 million records for books, compared to 1,278,000 in C19’s NSTC, and is often a better source for obscure titles. WorldCat also allows for the compilation and display of the many FRBR expressions of a given work by publication date, providing a quick, if incomplete, view of a book’s publication history. NSTC, however, remained the better source for dating popular works: it is far easier to locate the 1813 first edition of Pride and Prejudice among the 80 odd records in NSTC, for example, than it is from more than 4,000 records of expressions of the title in WorldCat.

2.2 Design choices
By grounding books in large-scale digital libraries in their earliest date of publication, we hope to enable more precise measurements of cultural phenomena that are realized in the language of the text; this fundamental use case drives several design choices in our manual dating process.

According to the FRBR model, the record for an expression of a given work (a translation, for example) would note both the publication date of a work in its original language as well as the date of the expression being cataloged. Since this has not been a consistent bibliographic practice, however, and because it goes beyond the scope of our interests here, we have chosen to annotate a single date for each item in our sample, following the guidelines below.

Translations. We have assigned translations that are the work of named translators in our sample with the earliest known date of the translation’s publication, not the publication date of the work in its original language. The Attic Nights, originally composed by Aulus Gellius sometime before his death in 180 AD, and translated into English by W. Beloe in 1795, is assigned the date of 1795, for example.

Serials. Portions of novels, as well as individual poems, short stories and essays, are often published in magazines and newspapers for years before their first publication as a unified work. We have dated the first printed editions of complete works rather than any individually printed components whenever there is clear evidence that these are author-initiated publications intended to appear as complete works.

Edited material. Popular works are often published and republished with prefatory materials, notes, and back matter composed at various times, and by a variety of authors and editors. We have chosen to select the date that best reflects the first publication of the core work, ignoring the presence of supplementary text.

2.3 Exclusions
There were several categories of books that could not be assigned any meaningful date of first publication, and so were not manually

dated and are thereby excluded from our sample. A substantial number of volumes defy the notion of encompassing a single work (compilations, for example). Others were poor candidates for large-scale text analysis, since they included too few textual elements or were in formats leading to poor OCR.

Compilations. Compilations may contain works composed or first published across a range of dates in non-serial publications, and therefore have not been dated in our sample. Included in this category are primary sources later collected and published: The Letters of Abelard and Eloisa (letters), An Account of the Life of the Reverend David Brainerd (diary entries), and Speeches, Messages, and Other Writings of the Hon. Albert G. Brown (speeches). Essays that are compiled by an editor or publisher are also included (e.g., Miscellaneous Essays, Volume Two of Modern British Essayists, by Archibald Alison). By far the most representative are collected works previously published in non-serial publications: e.g., The Dramatic Works of Beaumont and Fletcher (plays), Anton Chekhov’s The Kiss, and Other Stories (short stories), or Riverside Press’ 1883 The Complete Works of Nathaniel Hawthorne, Volume 3 (novels). Poems previously published over a wide range of dates (e.g., The Poetical Works of S.T. Coleridge) fall in this category while those released as a single volume by an author (e.g., Poems in Oil and Other Verse, by Will Ferrell), and book-length poems (e.g., The General by Francis Gentleman) do not. This category does not include collections of short stories for which we were not able to find clear evidence of their being previously published in non-serial titles (e.g., Stories of New England Life; or Leaves from the Tree Igdrasyl by Martha Russell). Compilations were found to be quite pervasive in all of our samples, accounting for approximately 20% of all books we attempted to date.

Abridgments. These texts are revised versions of another original work, often including much of the same text verbatim. We have chosen not to date these works since it is unclear how much of the original source is included. The Student’s Hume: A History of England by David Hume (1859), for example, “incorporates] the corrections and researches of recent historians,” but those corrections are undocumented in the text.

Folk tales. This category includes common stories for which there is no clear attribution of an individual author. George Routledge and Sons undated edition of Cinderella, for example, is only one of thousands of expressions of this tale.

Non-text. This category includes sheet music (e.g., Troubadour Song by Harriet Browne), tables (e.g., Observations on the State of the Air, Winds, Weather by Joseph Dymond), directories, lists, maps, dictionaries (e.g. Chinoook Vocabulary, Chinook-English), images, manuscript facsimiles (e.g., The World and the Child), or non-English languages. Approximately 5% of books in our collection were found to belong to this category.

3 ANALYSIS
After excluding books for dating using the criteria above, the final annotated dataset contains 916 manually dated books in the Uniform sample, 991 books in the Fiction sample, and 799 books in the Stratified sample. For the experiments reported below,
we hold out 500 books from each sample for evaluation and use the remainder for training and model selection (including hyper-parameter optimization).

<table>
<thead>
<tr>
<th>Sample</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>416</td>
<td>500</td>
</tr>
<tr>
<td>Fiction</td>
<td>491</td>
<td>500</td>
</tr>
<tr>
<td>Stratified</td>
<td>299</td>
<td>500</td>
</tr>
</tbody>
</table>

Table 1: Size of the different samples.

How much of a difference is there between the true dates of first publication and the publication dates of specific printings? Table 2 presents these results for all of the labeled data, along with 95% confidence intervals calculated using the bootstrap [9]; on average, the absolute difference among books in the uniform dataset is 2.73 years; among fiction books, the difference is approximately 4.66 years, and among those in our sample stratified by decade, the difference is approximately 5.09 years. While the mean combines the large majority of books with no (or minimal) distance along with outliers with an extremely large difference (e.g., several hundred years), we find that approximately 10% of books across our samples have a difference greater than 10 years; 1% of them have a difference greater than 100 years.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Mean absolute Δ</th>
<th>≥ 10</th>
<th>≥ 25</th>
<th>≥ 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>2.73 [2.1, 3.48]</td>
<td>7.1%</td>
<td>3.3%</td>
<td>0.3%</td>
</tr>
<tr>
<td>Fiction</td>
<td>4.66 [3.49, 6.12]</td>
<td>10.3%</td>
<td>5.5%</td>
<td>0.8%</td>
</tr>
<tr>
<td>Stratified</td>
<td>5.09 [4.01, 6.53]</td>
<td>10.8%</td>
<td>5.3%</td>
<td>0.9%</td>
</tr>
</tbody>
</table>

Table 2: Difference between the manually identified first publication date and that reported in the metadata, as measured by mean absolute error (with 95% confidence intervals), and percent of books with a difference of at least 10 years, 25 years and 100 years.

4 AUTOMATIC DATING

4.1 Metadata

The metadata provided in MARC records that attend books in the HathiTrust can be used to estimate the date of first publication in several ways. Date information in MARC records generally appears in the $008$ subfield (appearing as volume-specific $Date 1$ and $Date 2$ regions (whose semantics depend on the type of publication they appear with) and can also appear in custom $z30$ tags. The HathiTrust determines the publication date as part of a rights-determination algorithm. For reprints, the earlier of the two dates (captured in $Date 1$) is used; for works whose publication date is also the copyright date, that date is used; for all other works, the later of all observed dates (in any field) is used. Since this algorithm prioritizes the determination of rights in assigning dates, it is generally conservative, preferring the later dates observed. We term this below as HATHI DATE.

Without using dates for the determination of rights, we can adopt an alternative method for estimating the date of first publication by prioritizing the earliest dates mentioned in the $008$ or $z30$ fields. For non-series, we simply take the earliest date mentioned among all dates observed; for series, which often list the date of publication of the first volume in the series as $Date 1$, we take as our prediction the earliest date mentioned in the $z30$ field. This has the effect of biasing our estimate earlier. We term this below as EARLY.

4.2 Metadata-based deduplication

Outside of the individual metadata records for books in the collection, we can also leverage the resource of the collection as a whole. The HathiTrust, like many large-scale digital libraries, contains works aggregated from several university libraries [23]; each of these libraries may contain multiple printings and editions of a given expression for a work, and we can leverage this pattern of duplication to find better estimates of the first date of publication.

To do so, we define a metadata-based duplicate with a simple and reproducible heuristic: two metadata records are duplicates if the first 25 characters in both the title and author fields are the same. The restriction to the first 25 characters mitigates some variation as a function of the granularity of titles (e.g., $The life and adventures of Robinson Crusoe$ vs. $Of York, mariner$). For any given work, we assign its date of publication to be the earliest among its identified duplicates. We term this below as METADATA DEDUP.

4.3 Content-based deduplication

Metadata-based similarity can fail in several ways. First, requirements on the similarity of titles and authors may not identify as duplicates those works with slight variants (e.g., $The life and adventures of Robinson Crusoe$ vs. $Robinson Crusoe$). Second, it may not identify at all works whose metadata is erroneous, or whose works are found in collections or serials (e.g., $Great Expectations$ vs. vol. 7 of Charles Dickens’ Works).

Content-based deduplication provides an alternative that has been used extensively in digital libraries [42–44, 49, 50]. To identify duplicates based on the content, we represent books as a set of shingles comprised of hashed word trigrams, leveraging minhashing [2] to generate a compact 500-dimensional representation. We then identify near-duplicates in the entire 5.6 million book dataset using locality-sensitive hashing [21] on those minhash representations (placing each book into 250 distinct buckets, each identified using two sequential minhash features). Near-duplicates are defined as those with a jaccard similarity above some threshold $k$, and we set $k = 150$ using cross-validation on the training partition of the gold data described in section 2. We term this below as CONTENT DEDUP.

A third deduplication-based approach is to leverage both metadata deduplication and content deduplication in a single model, assigning the earliest date generated by either process. We term this COMBINED DEDUP below.
4.4 Results

Table 3 illustrates a comparison of the metadata- and deduplication-based methods. While the Early methodology on its own uniformly improves upon the dates of publication provided by the HathiTrust by their rights-determination algorithm for this particular purpose, we see that deduplication methods yield substantial improvement, leading to a 25-30% reduction in error over metadata alone. While Combined Deduplication performs the best on two of the samples, the simpler method of deduplicating based on the metadata alone performs best for the Uniform sample and is competitive across all three samples.

5 CONTENT-BASED PREDICTION

Metadata and deduplication-based methods of establishing the date of first publication depend on having either high-quality metadata or relatively large collections in which meaningful duplicates can be found. To explore the possibility of predicting the first date of publication from the content of the text itself, we assess two different approaches: one based on existing aggregate information from Google Books (in the context of a Naive Bayes classifier), and one based on training a discriminative model (linear regression with $\ell_2$ regularization) directly on a large sample of books from the HathiTrust.

For both models, we represent a book as a bag of unigrams, but only consider text that appears after the first 10% of pages at the start of the book (and a minimum of 10 pages), and before the final 6% of pages at the end of the book (which, in our data, were the average beginning and end of the core content of a book). This encourages the models to only have access to linguistic features of the text, and not any bibliographic data present in the frontmatter or general paratext (such as advertisements or publisher’s information at the end). This design choice allows us to assess the degree to which content-based prediction is complementary to metadata-based prediction (and not using similar information in making predictions).

5.1 Google Ngrams + Naive Bayes

One method for estimating a likely date of publication is to use existing aggregate resources from large-scale digital libraries, such as Google Books. The Google Ngram dataset contains ngram counts (unigrams up to 5-grams) for a total of 4.5 million books published between 1505 and 2008, in effect providing us with pre-calculated language models for each year in their collection (though the year of publication is that for specific editions, and not the date of first publication of any expression of the work).

With a Naive Bayes model, we make a categorical prediction for a year $y \in \mathcal{Y} = \{1700, \ldots, 1922\}$ given a set of observed words in a book $W$, our estimate of the prior belief over publication dates absent any data ($\theta$) and a set of unigram language models $\phi_i$ one for each year in the label space ($\phi_i \forall i \in 1700, \ldots, 1922$). Each $\phi_i$ lies in the $V$-dimensional probability simplex; in the experiments that follow, $V = 100,000$ (the 100,000 most frequent words in the data overall). Since Google Ngrams provides count data for terms by year, we simply set $\theta$ to be the empirical frequency of those terms as observed in that year, plus a small amount of additive smoothing, normalized by the total counts for that vector. Since our prior belief about the likely years differs according to our sample (e.g., our uniform sample and decade-stratified sample have different distributions of years by design), we estimate it empirically from the years observed in the training data.

This approach presents two issues: first, we are treating $y$ as a categorical variable rather than the ordinal one we know it to be. Second, the Google Ngram data is quite variable when aggregated at the level of any individual year, and reflects idiosyncrasies in the samples of which books are published (and how many of them).

As figure 2 illustrates, the raw relative frequency of the term *thee* has substantial fluctuation, especially among the less-well-attested years before 1800. This variability is such that the term is used twice as frequently in 1717 than in either 1716 or 1718.
that the unigram language models of adjacent time periods should also be similar (e.g., \( \hat{\phi}_{t-1} \approx \hat{\phi}_t \approx \hat{\phi}_{t+1} \)). To encode this assumption into our model, we convolve our empirical estimates of \( \hat{\phi} \) with a \( K \)-dimensional kernel \( k \), equivalent to calculating a moving average over \( \hat{\phi} \) along the time dimension, producing a new set of temporally smoothed unigram language models \( \hat{\phi}^s \).

\[
k = \left[ \frac{1}{K} \right]^K
\]

\[
\hat{\phi}^s = \hat{\phi} \ast k
\]

We then calculate the probability of year \( y \) for a book with words \( w \) through the usual application of Bayes’ rule:

\[
P(y | \theta, \hat{\phi}^s, w) \propto P(y | \theta) \prod_w P(w | \hat{\phi}^s_y)
\]

In this model, \( K \) (the convolution size) is a tunable parameter, and we optimize it on the training data.

### 5.2 HathiTrust + Ridge regression

Naive Bayes makes a strong simplifying assumption: that all features are independent and contribute equally to the prediction. This can be problematic in some scenarios where a large number of features are highly correlated with each other, leading to artificially high confidence in a prediction simply due to the fact of repeated but not particularly informative features.

In using an existing dataset, we may also be hampered by the corpus selection process; without knowing exactly how the Google Ngram counts were generated (and, in particular, which books they were generated from), we lose some control over our analysis. (In a very real possibility, the ngram counts could be generated from non-deduplicated data, where many editions of the same work or even copies of the same edition in different libraries, are treated as different books in generating the counts.)

To account for both of these factors, we also train a discriminative linear regression model with \( \ell_2 \) regularization (ridge regression) using books from the HathiTrust.

We train using three different datasets, each corresponding to one sampling strategy described above. The \textsc{Uniform} dataset is a sample of 100,000 books selected at random from all the works in the public domain of the HathiTrust, but—importantly—excluding all books written by the same author as any book in our test dataset (which thereby also excludes all of our test books from being included in the training sample). We treat all authors who share the same first 25 characters of their name as referring to the same individual. This exclusion yields a total dataset of 43,517 books.

The \textsc{Fiction} dataset includes all books in the Underwood 2014 [46] fiction dataset, with the same exclusion criteria as described above. This yields a total of 54,802 books.

The \textsc{Stratified} dataset includes a random sample of all books in the HathiTrust, but stratified by their date of publication. Using the same exclusion criteria yields a total of 14,547 books.

For all books in the training set, rather than using the observed date of publication as the true label whose prediction we are trying to optimize, we use the metadata deduplication strategy discussed in §4.2 to assign its label to be the date of publication for the earliest work among its identified duplicates. We feature each book as a set of binary indicators for all of words it contains that are also in the vocabulary of the 100,000 most frequent words overall, and again exclude all information that appears in the first 10% of last 6% of pages in the book.

### 5.3 Results

To compare the performance of the different models, we calculate the mean absolute error between the predictions they make on the test data \( \hat{y} \) and the manually identified gold labels for those books \( y \):

\[
\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i|
\]

In order to quantify our uncertainty around this measure given our sample, we calculate 95% confidence intervals using the non-parametric bootstrap [9], using 1000 bootstrap resamples of the test data and calculating the mean absolute error for each.

Table 4 presents the results of this analysis. In order to help contextualize the results, a simple baseline that predicts the average absolute value of all books in the training partition of the data yields a mean absolute error of 22.9 years for the \textsc{Uniform} dataset, 25.7 years for the \textsc{Fiction} dataset, and 46.3 years for the \textsc{Stratified} dataset. While ridge regression and Convolutional Naive Bayes are equivalent for the harder problem of predicting the date of publication for the \textsc{Stratified} dataset (for which there is both less training data and greater variance among the dates), ridge regression with binary features strongly outperforms the Naive Bayes models built from existing Google Ngram data for both the \textsc{Uniform} and \textsc{Fiction} samples, leading to a substantial reduction in error between 40-50%. Training a discriminative regressor directly on the full text of individual books, rather than relying on aggregate counts for entire years, leads to measurably stronger predictions.

### 6 ANALYSIS

As the automatic dating task above illustrates, the best method for estimating the date of first publication for books in a large digital library is to leverage the depth of the collection, identifying duplicates and assigning the first date of publication for a book to be the earliest date attested among its near-duplicates. While
Table 5: A sample of historical novels, along with their date of first publication, narrative time within the story, and predicted date according to our model.

<table>
<thead>
<tr>
<th>Author</th>
<th>Title</th>
<th>Narrative time</th>
<th>Date of publication</th>
<th>Predicted date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arnold Bennett</td>
<td>The Old Wives’ Tale</td>
<td>1872</td>
<td>1908</td>
<td>1914</td>
</tr>
<tr>
<td>John Galsworthy</td>
<td>The Man of Property</td>
<td>1870</td>
<td>1906</td>
<td>1907</td>
</tr>
<tr>
<td>Winston Churchill</td>
<td>The Crossing</td>
<td>1774</td>
<td>1904</td>
<td>1904</td>
</tr>
<tr>
<td>Stephen Crane</td>
<td>The Red Badge of Courage</td>
<td>1863</td>
<td>1895</td>
<td>1897</td>
</tr>
<tr>
<td>George Moore</td>
<td>Esther Waters</td>
<td>1875</td>
<td>1894</td>
<td>1897</td>
</tr>
<tr>
<td>Robert Louis Stevenson</td>
<td>The Master of Ballantrae</td>
<td>1745</td>
<td>1889</td>
<td>1878</td>
</tr>
<tr>
<td>Marcus Clarke</td>
<td>For the Term of His Natural Life</td>
<td>1827</td>
<td>1874</td>
<td>1879</td>
</tr>
<tr>
<td>Elizabeth Gaskell</td>
<td>Sylvia’s Lovers</td>
<td>1790</td>
<td>1863</td>
<td>1865</td>
</tr>
<tr>
<td>Charles Dickens</td>
<td>A Tale of Two Cities</td>
<td>1794</td>
<td>1859</td>
<td>1867</td>
</tr>
<tr>
<td>Walter Scott</td>
<td>The Bride of Lammermoor</td>
<td>1707</td>
<td>1830</td>
<td>1851</td>
</tr>
<tr>
<td>James Fenimore Cooper</td>
<td>Last of the Mohicans</td>
<td>1757</td>
<td>1826</td>
<td>1836</td>
</tr>
<tr>
<td>Walter Scott</td>
<td>The Heart of Midlothian</td>
<td>1736</td>
<td>1818</td>
<td>1840</td>
</tr>
<tr>
<td>Walter Scott</td>
<td>Rob Roy</td>
<td>1715</td>
<td>1817</td>
<td>1834</td>
</tr>
</tbody>
</table>

automatically predicting the date of first publication from the content of a book is not as precise, it can still serve several important ends. As Guo et al. (2015) [16] point out, a substantial fraction of books in the HathiTrust (13%) are missing publication date metadata, and even when present, information in legacy metadata can often be improved [13]; automatically estimating the date of publication, or an interval in which publication is likely, can help with the search and discovery of these texts when date is provided as a facet.

Another end is to analyze the degree to which the predicted date of composition agrees or disagrees with a given date of publication. Practically, books whose predicted and given dates strongly disagree may be good candidates for quality assurance. They can also provide the raw material for an analysis of what a content-based date prediction system is actually learning. When making predictions, are we relying more on historical markers in the text (e.g., mentions of historical figures like George Washington or explicit dates like 1776 that anchor a text in the late 18th century) or are we relying more on linguistic and stylistic signals characteristic of the time in which the book was written (such as thee and thou)?

To illustrate this, we consider a small selection of historical novels from the Fiction dataset, where there is a significant difference between the date of composition and the narrative time within the book; for example, James Fenimore Cooper’s Last of the Mohicans was written in 1826 but narrates a story that takes place in 1757. To identify historical novels, we leverage a list of Best Historical Fiction on GoodReads3 and identify the set intersection between that list and the books in our Fiction dataset, also adding the further constraints that the books must be written in English and describe a narrative time after 1700. Table 5 lists the 13 books that meet these criteria; this is small sample, and can best serve as an anecdotal case study of the potential for these methods.

To compare the predicted and given dates of publication for these historical novels, we train our best-performing model from table 4 above (ridge regression), in a ten-fold cross-validation of the full fiction dataset, taking care that works by the same author never show up across folds (so that we do not train on one copy of Great Expectations in one fold and use that information to estimate the date for another copy in a different fold).

In doing so, we use a model trained on 90% of the data to make predictions for the held-out 10%, and iterate through ten folds to make predictions for the entire dataset. As table 5 shows, we see that the predictions our models makes for these historical novels much more closely align with their actual date of first publication than the imagined historical time within each novel; the markers of linguistic style rather than the historical content appear to be driving the prediction decision.

Given the design of our model, we can also ask what the most indicative terms are within each book that lead to the estimate. In using linear regression, we make predictions for a book, represented as a \( V \)-dimensional feature vector \( x \in \mathbb{R}^V \), by calculating the dot product with a corresponding set of feature coefficients \( \beta \in \mathbb{R}^V \):

\[
y = \sum_{i=1}^{V} x_i \beta_i
\]

Martens and Provost (2014) [31] provide one method for explaining binary classification decisions in linear models by asking the following question: what is the minimal set of features that \( x \) has that, if removed, would lead us to predict the opposite label? We can extend this line of reasoning to the case of regression as well (where we predict a continuous variable rather than a discrete category) by asking: given a prediction \( \hat{y} \) for a given data point \( x \), what is the minimal set of features of \( x \) that we can remove in order to predict an alternative target \( \hat{y}' \)? To illustrate with one simple case study, what are the set of features we can remove from Last of the Mohicans to change our predicted date from 1835 to the narrative time of 1757?

As in Martens and Provost (2014), we can do so simply by ranking the learned coefficients \( \beta \) and removing features from \( x \) with the strongest positive weights (to push the prediction earlier in time) or the strongest negative weights (to push the prediction later in time). Removing 328 of Last of the Mohican’s 8976 features (3.7%) yields a target prediction \( \hat{y}' \) of 1757. The ten strongest
weighted features are shown in Table 6 and reveal the set of ahistorical terms that Cooper uses in the novel that are not only uncharacteristic of language spoken in 1757, but are also even on the leading edge of wider adoption, as we can confirm with an external data source in Google Ngrams. Figure 3 illustrates the inflection point between the use of *every one* and *everyone*, and Figure 4 illustrates the precipitous rise of *later*, both of which Cooper is an early adopter (which also in part leads to a higher predicted date of publication of 1835 than its true date of first publication of 1826).

<table>
<thead>
<tr>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>etc.</td>
</tr>
<tr>
<td>later</td>
</tr>
<tr>
<td>everything</td>
</tr>
<tr>
<td>everybody</td>
</tr>
<tr>
<td>anything</td>
</tr>
<tr>
<td>ahead</td>
</tr>
<tr>
<td>lack</td>
</tr>
<tr>
<td>big</td>
</tr>
<tr>
<td>simply</td>
</tr>
<tr>
<td>meantime</td>
</tr>
</tbody>
</table>

Table 6: Ahistorical terms in James Fenimore Cooper’s *Last of the Mohicans*.

Even though our estimates of the true first date of publication are better served with deduplication-based methods, learning a model to predict this date from the content of the book gives us the potential for deeper insight into the books in our collection by providing a mechanism for measuring apparent time, as distinct from both the observed publication date or the narrative time.

![Figure 3: Relative frequency of *everyone* and *every one* in the Google Ngram data.](image)

![Figure 4: Relative frequency of *later* in the Google Ngram data.](image)

7 RELATED WORK

Our approach and general methodology here are motivated and informed by much prior work in automatically predicting the date of publication for text, including newspaper articles [41], web pages [24, 39], Wikipedia biographies [28], literature (such as short stories from Project Gutenberg [29] and Romanian novels) [4], and large digital libraries [16, 30].

Nearly all of these approaches choose an experimental design used by de Jong et al. (2005) [8], which partitions time into discrete buckets (e.g., 1870-1880) and attempts to classify a new document into one of those buckets—rather than, as in our case, attempting to pinpoint the exact date of publication (as a regression problem). There are tradeoffs to both design choices: classification into discrete buckets allows for the natural incorporation of features based on the distribution of a language model expressed in that bucket [8, 16, 24, 29] at the cost of the specificity of the prediction (a model may be able to predict with finer detail than the bucket allows); regression is more limited in the information that can be brought to bear on the problem but potentially allows for finer-grained predictions, while also enabling the analysis discussed in section 6: understanding what features contribute to a prediction, and isolating the specific characteristics of a text that give its date away.

While other work has also focused on the prediction of publication dates in large digital libraries (either the Hathi-Trust [16] or Google Books [30]), one further contribution we make is distinguishing between the different meanings of “publication date” when seen in the context of the FRBR hierarchy: all work to date has leveraged the publication date provided in the metadata for training and evaluation, effectively focusing on learning the temporal characteristics at the FRBR level of a *manifestation*—when a specific edition was published. This variety of temporal metadata certainly has its use cases, but is only one choice among several: in creating a dataset and evaluating our predictions on the first date of publication, we are effectively assessing our ability to learn the temporal characteristics of books at the FRBR level of a work.
Related to work in the explicit prediction of the publication dates of text is adjacent work in estimating the change in the meanings of words over time [14, 17, 22, 26, 27, 33, 35]; while the goal of this work is often to characterize the linguistic dynamics of change, and to estimate when several different words across time all refer to the same (or similar) concept, it can also be used in the design of features that provide information not simply on the presence, absence or frequency of a specific lexical form (as in this and other past work), but also when a given lexical form is used in a way that is particularly characteristic of specific time periods (e.g., while *apple* is a common word throughout the history of print, it only recently is used in the context of a specific organization).

In characterizing the temporal signature of time periods and providing a way to measure the degree to which an author or a specific book predates the wider adoption of some linguistic phenomenon, one potential future direction for this work is in analyzing the diffusion of stylistic innovation, especially in the creating and popularization of neologisms [3, 5, 6, 34].

8 CONCLUSION

We present in this work a new gold standard dataset and several empirical analyses for predicting the first date of publication for books in a large-scale digital library, making an important distinction between the date a specific FRBR manifestation was published and the date its original work (of which it is an instantiation) was published.

As more and more work in the digital humanities, computational social science and cultural analytics is increasingly making use of the texts and their attendant metadata in large-scale digital libraries, this is an important distinction to be made; both measures are appropriate for different analyses, and, as we show, a simple metadata-based deduplication method is often acceptable for estimating the work publication from a set of manifestation dates in a large enough collection. When metadata is lacking, content-based estimation can also yield relatively accurate measures, and can itself occasion analyses on the linguistic and stylistic characteristics of books and the authors who write them.

All annotated data is available for public use under a Creative Commons license at: https://github.com/dbamman/jcdl2017.

9 ACKNOWLEDGMENTS

We thank the anonymous reviewers and Ted Underwood for valuable advice, and are grateful to the Hathitrust Research Center for their assistance in enabling this work. The research reported in this article was supported by a grant from the Digital Humanities at Berkeley initiative.

REFERENCES


