A machine-learning approach to coding book reviews as quality indicators: towards a theory of ‘mega-citation’

Alesia Zuccala*, Maarten van Someren and Maurits van Bellen**

*a.a.zuccala@uva.nl (corresponding author)
Institute for Logic, Language and Computation, Faculty of Humanities, University of Amsterdam,
Science Park 105, Amsterdam, 1090 GE (The Netherlands)

**m.w.vanSomeren@uva.nl; mauritsvanbellen@gmail.com
Informatics Institute, Faculty of Science, University of Amsterdam
Science Park 107, 1098 XG Amsterdam (The Netherlands)

Abstract

A theory of 'mega-citation' is introduced and used in an experiment to demonstrate how a qualitative scholarly book review can be converted into a weighted bibliometric indicator. We employ a manual human-coding approach to classify book reviews in the field of history based on reviewers' assessments of a book author's scholarly credibility (SC) and writing style (WS). A total of 100 book reviews were selected from the American Historical Review and coded for their positive/negative valence on these two dimensions. Most were coded as positive (68% for SC and 47% for WS respectively) and there was also a small positive correlation between SC and WS (r= 0.2). We then constructed a classifier, combining both manual design and machine learning, to categorize sentiment-based sentences in history book reviews. The machine classifier produced a matched accuracy (matched to the human coding) of approximately 75% for SC and 64% for WS. Writing style (WS) was found to be more difficult to classify by machine than scholarly credibility (SC) due to the use of subtler language of many reviewers. With further training data a machine-learning approach could be useful for automatically classifying a large number of history book reviews at once. Weighted 'mega-citations' can be especially valuable if they are used in conjunction with regular book/journal citations, and ' libcitations' (i.e., library holding counts) for a comprehensive assessment of a book/monograph's scholarly impact.

1. Introduction

A scholarly book review is a rich, though sometimes subtle source of information about the perceived value of a newly published book or monograph. In the past, book reviews have been examined on the basis of their structure and rhetorical content (Hartley, 2010; Motta-Roth, 1998; Nicolaïsen, 2002c), their usefulness to librarians for developing book collections (Blake, 1989; Serebnick, 1992), their importance to academic research and teaching (Hartley, 2006; Spink et al, 1998), and ‘citedness’ in the journal literature (Diodato, 1984; Zuccala & van Leeuwen, 2011). Here, we consider a new role for the book review as a ‘mega-citation’, or, quantitative quality-based measure for use in research evaluation procedures in fields across the humanities.
Our objective is to experiment with a machine-learning approach to classifying book reviews (‘mega-citations’) published in a scholarly review journal. This approach involves training an intelligent system to recognize parts of a review where a review writer has positively or negatively evaluated a book based on the book author’s scholarly credibility and writing style. We will compare the systematic coding with the eye of a human to the coding that we can achieve by training an intelligent computing system.

Book review texts – 100 in total – were selected from the American Historical Review (i.e., February issues from the years 2000, 2003, 2005, and 2007). To facilitate the analyses we chose a set of book reviews that were relatively uniform in style and of considerable length compared to some written in other fields. Reviews published in this journal (200 per issue) are normally between 2,000 to 3,000 words. The American Historical Review (2012) editorial board does not “dictate the content of reviews”; however, persons who serve as reviewers for this journal are “expected to have earned a Ph.D.” and are required to adhere to a certain academic standard.

This experiment focuses only on the subject of history, because book reviewing is important to this field of scholarship, and reviews constitute a significant portion of the documents regularly published in history journals (Bilhartz, 1984; Stowe, 1991; Zuccala & van Leeuwen, 2011). Our aim is to illustrate how book reviews as ‘mega-citations’ may be used in connection with regular citations (i.e., cited references appearing in journal research articles, or other books) and ‘libcitations’ or library catalogue inclusions (White et al., 2009; Torres-Salinas & Moed, 2009) to measure a history book’s overall quality, scholarly influence and perceived cultural influence. Bibliometricians have spent comparatively more time evaluating science journals as opposed to books published in the humanities and other forms of humanistic output (Haustein, 2013); however, there is a critical need within this latter field to develop meaningful and robust research indicators and to explore new procedures for evaluation (Guillory, 2005).

2. Literature review

2.1 Some background theory on citations

One of the earliest contributions to citation theory was a paper written by Kaplan (1965), who argued that citations bestowed a certain "intellectual and scientific respectability on citing papers (p. 182). Since then, the theoretical debate has taken many turns, and for the most part it has remained both sociological and scientometric in nature (Leydesdorff, 1998; Wouters, 1999). On the one hand we have the normative perspective of citation, and on the other, a rhetorical view. Do we cite a research article because it is the ‘norm’ to give recognition to our peers in the scientific reward system (Merton, 1996), or, do we cite a piece of research to substantiate an argument and persuade the reader, or both (Cozzens, 1989; Gilbert, 1977)? Some scholars appreciate citations as physical entities where ensembles of citers can be analyzed according to a thermodynamic theory of citing (Price, 1980; Van Raan, 1998), while others think of citations in symbolic terms (Garfield, 1964; Small, 1978). In addition, co-citations in reference lists can be ‘mapped’ to show a history of influence within a discipline (White & Griffith, 1981). Underlying the citation as object or symbol there are also reasons for citing; that is, socially defined notions
of quality within a scholarly community, including the motivations of citing authors themselves (Case & Higgens, 2000; Cole & Cole, 1967; Cronin, 1984; Leydesdorff & Amstersamska, 1990).

Extensive reviews have been written about these theories (Bornmann & Daniel, 2008; Cozzens, 1981; Cronin, 1998; Moed, 2005; Wouters, 1999), but so far, no one citation theory has really superseded another as being primary or all encompassing, since each one has its proper place (Wouters, 1999). Moed (2005) also indicates that “extreme positions are not useful” in the overall theoretical debate. For instance, it is difficult to prove consistently that a collection of cited references do not possess any thematic links or elements in common (i.e., the constructivist view), or to assert that a citation is in fact the only valid measure of research quality (i.e., the citationist view) (see Moed, 2005, p. 213).

Amid this great debate we assume that the monograph plays a role, but in research focused on citations to books or within books, less emphasis is placed on theory and more on basic citation monitoring. Researchers are inclined to count and/or classify citations to and from monographs in the usual manner akin to journal articles (Cronin & Snyder, 1997; Cullars, 1992; 1998; Frost, 1979; Knievel & Kellsey, 2005; Kousha et al., 2011; Lewison, 2001; Nederhof, 2006). Two notable exceptions include another paper written by Torres-Salinas and Moed (2009) and another written by White et al. (2009). Both explore the potential of library catalogues as tools for bibliometric analyses, where an analogy is created between journal-based citations and library holdings. Torres-Salinas and Moed (2009) focus their study on the number of catalogue inclusions per book title in WorldCat®, while White et al. (2009) introduce the term ‘libcitation’, which may be taken as “an indicator of perceived cultural benefit” (p. 1087).

The advent of Thomson Reuters’ (2012) Book Citation Index (BCI) implies that opportunities will soon be available to develop bibliometric performance indicators for books (Thomson Reuters, 2012). Torres-Salinas et al. (2012) recently considered how indicators might be conceived and calculated from this new index, focusing specifically on book publishers. We also suggest that it is time for a fresh view of indicators and citations; however, our interest lies with the reviews that have been written about books, and not exclusively with the books themselves. In its basic form, a book review is a type of publication designed to give citation(s), but not necessarily to receive them. Book reviews are becoming more ‘scholarly’ in terms of citing documents in addition to the book under review (Nicolaisen, 2002b), but it is true that they are rarely cited (Diodato, 1984; Zuccala & van Leeuwen, 2011). This means that the book review has a special role within the scholarly communication system, but how can it be defined? Is it simply a short, evaluative, public essay or is it actually a special type of citation, or both?

2.2 The book review as 'mega-citation'

The role of a book review is to alert scholars to the value of a newly published book. Our view is that it functions as a kind of grand citation or megacitation devised to indicate the quality of that book. Normally the ‘mega-citation’ is a short piece of approximately 600 to 3000 words, written about one book/monograph published by a single or multiple authors. When a scholar is invited to prepare a review (usually scholars are asked), there is a basic requirement to ‘cite’ the book’s author(s), title and publisher. A professional review itself is then organized on the basis of
rhetorical moves and sub-functions (Motta-Roth, 1998). Both Nicolaisen (2002c) and Hartley (2010) agree with this notion (i.e., both also citing Motta-Roth, 1998), but Hartley (2010) explains that "language used in such reviews can become a code: what is said is not always what is meant" (p. 477). In other words, if a reviewer says: "this is a blockbuster of a text", he or she probably means that it is "enormous" or perhaps "too enormous" for the reader's liking (p. 477).

With the first rhetorical move, the reviewer usually introduces the book and provides details regarding the topic, the author(s), and the field in which the book is situated. The second ‘move’ is to present an outline of the book. In the third rhetorical move, the reviewer highlights specific parts for appraisal, commenting on both good and bad elements; while the final paragraph is normally reserved for an overall evaluation - i.e., one last move towards accepting or disqualifying the book in positive or negative terms (Motta-Roth, 1998). Normally the reviewer and the author of the monograph are peers from the same discipline, but not necessarily. Lindholm-Romantschuk (1998) has found that a significant number of book reviews published in one type of disciplinary journal are reviews of books originating in other disciplines. Clearly, they have potential to introduce books to scholars from other fields and provide insight into topics that researchers otherwise might not consider reading or using to further their knowledge.

In the reward system of science, the author of a book review accrues little to no benefits for his/her work. With respect to ‘giving credit where credit is due’, scholars surveyed from the natural sciences, social sciences and the humanities tend to agree that it would be good to receive institutional recognition for writing reviews (Hartley, 2006). Other aspects of academic work are similar to writing a book review, like refereeing journal articles for publication; however, the latter does not result in a formal, published piece read by the scholarly community at large. A book review is mega because it is rich in content, and because there is evidence that as a megacitation it possess some influence and predictive power: books that are reviewed positively tend to receive more citations in journal articles, than books that have been evaluated in neutral or negative terms (Nicolaisen, 2002a; Schubert et al., 1984).

2.3 Toward a theory of mega-citation

According to Holton (1978), new bibliometric indicators “cannot be thought of as given from ‘above’, or detached from the theoretical framework” and “should preferably be developed in response to and as aids in the solution of interesting questions and problems” (p. 55). At present, we see a problem. We do not have a bibliometric approach to indicating the quality of a book, which takes into account book reviews. This is significant due to the fact that books and their reviews tend to be important modes of output, particularly for disciplines like history and literary theory/criticism (Zuccala & van Leeuwen, 2011). The process of communicating knowledge in many fields across the arts and humanities is not typically formed by journal publications but rather books/monographs: “monographs are like the main course of a meal, journal articles and other scholarly communication are like tapas” (Williams et al., 2009, p.76).

To develop our theory, we start with the following questions: What does the book review as ‘mega-citation’ measure and how does it differ from other citations? Insights may be borrowed from the ‘contextualist’ school of citation, where a connection between citing and cited text is based on the function of the cited reference in the citing text’s argumentation structure. (Chubin
A journal article is constructed on the basis of a many-to-one relationship: many texts (i.e., articles; books; proceedings articles; etc.) may be cited to build one piece of work. Every external piece of information appearing in the article varies and can be "cited at different levels of granularity or aggregation" (Cronin, 1994, p. 537). For example, a typology of tiered citation starts at the level of the *oeuvre* (e.g., a collection of works by an author), progressing further to include a *motif*, an *opus*, a *chunk*, or a *quantum* of information from the text such as a formula, algorithm, phrase, name, etc. (Cronin, 1994). Moravcsik and Murugesan (1975) do not think in terms of tiers, but citation taxonomies: *conceptual* versus *operational*, *organic* versus *perfunctory*, *evolutionary* versus *juxtapositional*, and *confirmative* versus *negational* (p. 88). And, in the work of Maricic et al. (1998) the type of citation observed within a document is said to be meaningful depending on the position it occupies – i.e., a citation location parameter.

![Diagram of Journal Article and Book Review](image)

Figure 1. Relationship between citing and cited text: many-to-one or parts-to-one.

With a book review, the citing-cited connection is characteristically different from the citation in that many parts of one document are used to build one review (see Figure 1). In theoretical terms, Pichappan (1996) refers to this as an *Umbral-Umbral* relationship, which means that the crux of the book review and the book is similar. A secondary *Umbral-Penumbral* relationship might occur if the reviewer cites one or two documents (usually another book) other than the book under review. In this case, the additional piece may be a smaller part of the discussion, but still "related to the core of the citing [review]" (Pichappan, 1996, p. 651).

The many-parts-to-one notion implies that the content of the review, like the citation proper, is a suitable element for in-depth study, and indeed, content analyses of book reviews have been carried out as early as the 1970s (Champion & Morris, 1973; Riley & Spreitzer, 1970; Snizek & Fuhrman, 1979). During this period Glenn (1978) claimed that reviews were "not as adequate for evaluating books and authors as many people seem[ed] to think" (p. 254). Bilhartz' (1984) manual analyses of 560 history book reviews (1950 to 1980) was somewhat indicative of this problem; however differences in review habits have evolved over time:
the reviewers of the fifties were less concerned with the scholastic contribution of the work and more concerned with its readability and narrative excitement... On the whole, this was an era for friendly and uncritical evaluation.

In the early sixties, the era of the 'gentlemanly reviews' came to an end. Reviewers increasingly devoted more space to critiquing rather than simply summarizing the book's content.

By the 1970s, the historian clearly had entered the world of the social scientist... Boldness of interpretation was rewarded [and] to a greater degree than before; reviewers took a strong interest in originality of method.

The trends of the seventies continued into the eighties with minor variations. While originality in method remained important, the early glamour of quantification faded... Sophisticated reviewers [were] as apt to criticize the misuse of statistical procedures as they [were] to applaud them. More than in any previous decade, reviewers [of the 1980s] expect[ed] histories to have a sharply focused and well-analyzed thesis (pp. 527-528).

By the early nineties, Stowe (1991) found that reviews had evolved from being considered quite lowly to becoming “deceptively simple, hiding a rich mixture of interpretation, assumption, and insight (p. 592).

Librarians were among the first to recognize book reviews as insightful aids for developing library collections (Blake, 1989; Dilevko et al., 2006; Natowitz & Wheeler Carlo, 1997; Parker, 1989; Serebnick, 1992); giving consideration also to the relationship between the review and the reputation of a book's publisher (Jordy et al., 1999). It is only recently that we have become interested in the book review's academic value and scholarly impact (Hartley, 2006; Nicolaisen, 2002b; Spink et al, 1998). Today we expect reviews published in journals to be scholarly; hence, the features that academics look for in a good book review include a well-known person as the review author, the presentation of a straightforward overview of the book, a strong critique of the book’s main argument, and a strong evaluation of the book’s academic credibility (Hartley, 2006; Miranda, 1996).

2.4 Mega-citations en masse

The citation ‘proper’ (i.e., observed for many years in journal articles) is said to operate within two systems: a rhetorical (cognitive) system and a reward system, or system of social relations between scientists (Cozzens, 1989). The practice of carrying out citation analyses became popular when scholars realized that citations were particularly useful as indicators of links between the social and the cognitive dimensions of science (Cronin, 1984; Garfield, 1983). For example, “the number of times an article was cited could be seen as an indicator of the performance of the cited author(s), and thus a translation was made from the cognitive use of citations in a text to the social system of rewards operating in the scientific community” (Leydesdorff & Amstersdamska, 1990, p. 307).
Book reviews as 'mega-citations' also form a link between the social and cognitive system of scholarship (see Table 1). An author of a monograph is rewarded socially by a review of his or her monograph shortly after it is published. The nature of this reward depends to some extent on the discursive nature of the reviewing text and its subtle evaluation of the reviewed text. The reviewing author serves as a cognitive resource for the reviewed text; hence it is important to know the professional relationship between who is reviewing and who has been reviewed (e.g., Are both from the same field? Are both academically accomplished and well known? Are they known to each other? Has one cited the other and vice versa?). Links between the professional status of the reviewer and the evaluative content of reviews have been studied, and in Taylor's (1967) doctoral research, lower-status reviewers were found to write more favorable reviews, especially of books written by higher status authors. Hirsch et al. (1974) later found the opposite: lower-status reviewers were more critical of higher-status authors.

Table 1. Links between the social and cognitive system of science for the book review.

<table>
<thead>
<tr>
<th>Reviewed Author</th>
<th>Reviewing Author</th>
<th>Reviewing Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reviewed Text</td>
<td>Professional relation</td>
<td>Reward</td>
</tr>
<tr>
<td></td>
<td>Cognitive resource</td>
<td>Discursive relation</td>
</tr>
</tbody>
</table>

En masse, journal citations are symbols of recognition and influence. When bibliometricians use them collectively it is not common to investigate the underlying motive of every citer or the discursive relation between every citing and cited text. Book reviews, on the other hand, tend to exhibit more sentiment. With the term ‘sentiment’ we mean that the review is likely to contain information in the form of an opinion, judgement, valuation or attitude about the reviewed work or author. Carlo and Natowitz (1996), for example, analysed how words of praise were used in Choice reviews written for outstanding book titles. While book reviews appear between one to three years after a book has been published (or sometimes not at all, depending on the publisher and book), citations can continue to appear indefinitely. It is for this reason that we might work with reviews a little differently, by distilling parts of their content for bibliometric evaluation purposes.

In the field of history alone, approximately 15,000 scholarly book reviews are published every year\(^1\). Manually coding reviewer sentiments present in hundreds or thousands of scholarly reviews is not a practical way to collect data needed for a bibliometric analysis. This requires automatic coding by a computer system. It may be possible to program a computer to perform this task, but it is very difficult. We can specify some cues that suggest high or low levels of sentiment detection, but most likely this will not lead to an effective program. We therefore take a machine learning approach, and according to Sebastiani (2002) this can achieve a level of accuracy comparable to that of a human coder. Prabowo and Thelwall (2009) have found,

\(^1\) This average count was taken from a search in the Web of Knowledge Arts & Humanities Index for book reviews published in history journals (“WC” category = “History”) over five years up to and including 2012.
however, that in the absence of a superlative classifier, multiple classifiers used in a hybrid manner can often improve the effectiveness of sentiment analysis.

A recent survey written by Tang et al. (2009) confirms that review texts are commonly used in sentiment detection analysis. For instance, Turney (2002) has evaluated the reviews posted on Epinions (www.epinions.com) and many other scholars have studied movie reviews from the Internet Movie Database (IMDB) (Mullen & Collier, 2004; Pang et al., 2002, Salvetti, et al., 2004). Machine-based sentiment detection algorithms are also useful for mining opinions from social network sites, blogs, and discussion forums (Thelwall et al., 2010). For the present study, we would like to propose a new framework for detecting sentiments from a corpus of scholarly book reviews.

3. Analytical framework

All book reviews as mega-citations may be said to evaluate two general elements: the scholarly credibility of a newly published book and the author's writing style. A specific review, or mega-citation can occupy a point on a two-dimensional credibility-style scale, based on sentences or phrases appearing in the review text. For instance, the reviewer's use of phrases such as “this analysis seems flawed...” or “[The author] gleans answers from a forest of detail” can indicate the reviewer’s assessment of a book author’s scholarly credibility. Likewise, other phrases or terms used by the reviewer, such as incisive, well written, prosaic, or poorly organised might indicate positive-to-negative assessments of writing style.

In the next section we examine how an intelligent system may be used to identify the rhetorical language of book reviews, that is, the way in which reviewers apply words or phrases in a review to indicate nuances of critique and appreciation.

4. Research Methods

Our approach to evaluating book reviews consisted of four steps. First, we tested our credibility-style framework by collecting and manually coding a corpus of book reviews. Second, we developed a system for classifying sentences in book reviews that combines both manual modeling and machine learning. We then evaluated this process on the coded sentences and reviews.

4.1 Collecting and manually coding reviews

The dataset used for our study was a selection of 100 book reviews published in the American Historical Review (AHR) – i.e., February issues from the years 2000, 2003, 2005 and 2007. The reviews were retrieved digitally in PDF format and then converted to plain text. One person with a significant scholarly background, who was not a historian, performed the coding task and all authors of this paper examined and agreed with the results. The coding itself involved locating the sentences in each individual book review pertaining to judgments of scholarly credibility and noting if the reviewer’s sentiment about this was either positive or negative (SC + / SC-). The
same approach was then used to identify and code negative or positive sentiments pertaining to writing style (WS + / WS -). Figure 2 illustrates this procedure with a fragment of one review written by Jorge Canizares-Esguerra, for the Patricia Seed's book titled *American Pentimento: The Invention of Indians and the Pursuit of Riches* (Minneapolis: University of Minnesota Press) (see Figure 2).

For each review text, and for both scholarly credibility and writing style, the number of negative comments made by a reviewer was subtracted from those found to be positive, then divided by the total number of sentences in the full text. The formulae, written below, specify how the scholarly credibility and writing style values were calculated. Figure 3, below, illustrates how each book review fits on a scatterplot of our two-dimensional credibility-style framework (see Figure 3).

![Figure 2. Manual sentence coding of scholarly credibility/writing style.](image)

**General formulae:**

\[
\text{ScholarlyCredibility (SC)} = 0.5 \left( \frac{SC_{\text{positive}} - SC_{\text{negative}}}{\text{AllSentences}} \right)
\]

\[
\text{WritingStyle (WS)} = 0.5 \left( \frac{WS_{\text{positive}} - WS_{\text{negative}}}{\text{AllSentences}} \right)
\]

*Where, SC + WS = WeightedScore*

Review (metric) weight for Jorge Canizares-Esguerra review of *American Pentimento: The Invention of Indians and the Pursuit of Riches*:
ScholarlyCredibility (SC) = 0.5 \left( \frac{4-1}{26} \right) = 0.5 \times (0.115) = 0.06

Writing Style (WS) = 0.5 \left( \frac{3-0}{26} \right) = 0.5 \times (0.115) = 0.06

Where, \( SC + WS = 0.06 + 0.06 = 0.12 \)

Figure 3 illustrates the variance in the opinions expressed by the AHR reviewers. While some reviewers praised the scholarly credibility of the book's author, others were in fact quite critical. We found many comments pertaining to the book author's writing style, though these were not as frequent as those pertaining to scholarly credibility. However, positive and negative sentiments for both dimensions tended to correspond and there was a small positive correlation (Pearson's product moment correlation \( r = 0.2 \)). The average number of sentences for each coded review was 28. Table 2 presents a sample of 8 coded reviews (from 100) and for each category the percentage of full reviews classified as positive, negative or neutral are presented in Table 3.

### Table 2. Sample coding results from 8 reviews out of 100 AHR book reviews.

<table>
<thead>
<tr>
<th>Year</th>
<th>Reviewer</th>
<th>Total Sentences</th>
<th>WS+</th>
<th>WS-</th>
<th>SC+</th>
<th>SC-</th>
<th>Writing Style</th>
<th>Scholarly Credibility</th>
<th>Writing Style (Judgement)</th>
<th>Scholarly Credibility (Judgement)</th>
<th>Writing Style (Scatterplot X-value)</th>
<th>Scholarly Credibility (Scatterplot Y-value)</th>
<th>Bibliometric Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>Cave</td>
<td>21</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>1.5</td>
<td>Positive</td>
<td>Positive</td>
<td>0.07</td>
<td>0.07</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>Conforti</td>
<td>36</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>0.5</td>
<td>Positive</td>
<td>Neutral</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>Zelzer</td>
<td>33</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>7</td>
<td>-1</td>
<td>Positive</td>
<td>Negative</td>
<td>0.06</td>
<td>-0.03</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>Armstrong</td>
<td>24</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>Neutral</td>
<td>Positive</td>
<td>0.00</td>
<td>0.06</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>Lavere</td>
<td>32</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>Neutral</td>
<td>Neutral</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>Breen</td>
<td>28</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>Neutral</td>
<td>Negative</td>
<td>0.00</td>
<td>-0.04</td>
<td>-0.04</td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>Stuart</td>
<td>35</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>-1</td>
<td>Negative</td>
<td>Positive</td>
<td>-0.03</td>
<td>0.07</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>Tilechim</td>
<td>31</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>11</td>
<td>-1</td>
<td>Negative</td>
<td>Negative</td>
<td>-0.03</td>
<td>-0.16</td>
<td>-0.19</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3: Percentage of full reviews classified as positive, neutral or negative.

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Neutral</th>
<th>Negative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC</td>
<td>68%</td>
<td>6%</td>
<td>26%</td>
<td>100</td>
</tr>
<tr>
<td>WS</td>
<td>47%</td>
<td>38%</td>
<td>15%</td>
<td>100</td>
</tr>
</tbody>
</table>
Figure 3: Two-dimensional scaling of 100 AHR book reviews.

4.2 Constructing a review sentence classifier

For this part of the study we used the data collected for the initial manual coding procedure to construct a text classifier, designed to categorize sentences in reviews by their scholarly credibility and writing style. One approach to constructing this classifier is to manually design and implement a program and a second approach is to collect sentences from reviews, label them and use machine-learning algorithms that will automatically construct a classifier from labelled data. Sometimes the manual programming approach is not feasible if there is a broad variety of language used in the texts. Background information on the use of machine learning for texts can be found in Feldman and Sanger (2006) or Weiss et al. (2010). In this case we were working with two aspects of language, opinions versus non-opinions, and positive versus negative opinions; hence our solution was as follows:

1. Manually build a classifier that recognizes sentences that express an opinion about SC or WS.
2. Use Machine Learning to construct an additional classifier that will decide whether these sentences are positive or negative for SC or WS.
The classifiers do not operate directly on text files. First the text files are pre-processed to remove irrelevant information and to use linguistic techniques to reduce the variation in the words. Reducing the variation simplifies the task for the classifiers (and the learning algorithm). Pre-processing consists of the following steps:

1. **Segmenting** the review into sentences. This is done by using a “.” followed by a space or a “new line” and a capital letter as separation between sentences.
2. **Tokenizing.** This is done using a standard tokenizer. A tokenizer splits a series of symbols into tokens. The tokens are the units for further analysis. For example, we may see “New York” as one token or as two tokens “new” and “york”. Another example from Figure 2: “wide-angled” can be seen as one or two tokens. “State University of New York, Buffalo” can be seen as a single token, or as two (“State University of New York” and “Buffalo”), or three (”State University”, (of) “New York” and “Buffalo”) or more. Another example from Figure 2 is “Seed’s”. This can be a single token or a pair (“Seed” and “s”). Relatively simple tokenizers tend to do a good job even if they make tokens that are suboptimal from a semantic viewpoint because they do so consistently for all documents under consideration. The one that we used can be found at https://lucene.apache.org/core/old_versioned_docs/versions/3_0_0/api/core/org/apache/lucene/analysis/standard/StandardTokenizer.html
3. **Removing** highly frequent words: Words with a very high frequency are removed to reduce the number of features. These words usually are not associated with class membership of texts. This is done using a list that is publicly available at http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop
4. **Stemming:** Words are replaced by their stem using the Porter stemmer (see http://tartarus.org/martin/PorterStemmer/). E.g., work, working, works and worked are all treated as being the same word.

### 4.3. Constructing a classifier for relevant review sentences

It is possible to apply the machine learning approach to the review texts as a whole but within the texts few sentences are used by the reviewer to express an opinion about the scholarly credibility or writing style of an author of a book. Normally, much of the review is dedicated to summarizing parts of a book’s content and describing a particular historical context.

Upon inspecting the reviews, we found that in most cases the sentences that were most relevant to a judgement of SC or WQ included words that refer to *people, research, argumentation* and *work* (i.e., *the book as a piece of written work*). We therefore constructed a domain ontology designed to collect words that fit within one of four prescribed categories (see Figure 4).
We then built a classifier based on the relative frequency of words from each of the four categories that appear in a sentence. Specifically, we measure if the frequency of tokenized and stemmed words is higher in the sentence than in the entire document set. This is used to classify a sentence as relevant because it expresses an opinion.

4.4 Classifying relevant sentences as positive or negative for SC or WS.

With the previous step, we were able to extract from each review all sentences that express an opinion about SC/WS and all of the useful words from each sentence. Classification of a sentence is based on the set of tokens that occurs in a sentence, without considering their order within that particular sentence. A sentence can contain an opinion on scholarly credibility and/or writing style and therefore each sentence is classified twice, both for SC and WS.

For the classification process we used a machine learning approach based on a combination of two methods, AdaBoost and the Naïve Bayes Classifier. Details of these methods can be found in any textbook on Machine Learning and the software that implements this method is also widely available in the public domain. The Naïve Bayes Classifier is based on Bayes Rule from probability theory, which states that:

\[ P(\text{Class} | \text{Words}) = P(\text{Class}) \cdot P(\text{Words} | \text{Class}) / P(\text{Words}) \]

The assumption of the Naïve Bayes Classifier is that \( P(\text{Words} | \text{Class}) \) is equal to:

\[ P(\text{Word}_1 | \text{Class}) \cdot P(\text{Word}_2 | \text{Class}) \cdot P(\text{Word}_3 | \text{Class}) ... \cdot P(\text{Word}_n | \text{Class}) \]

For all words and classes we basically assume that within a Class the Words are independent. This assumption is of course never true; nevertheless the Naïve Bayes Classifier is still useful in terms of producing a good result.

To classify a sentence we need \( P(\text{Class}) \), the “prior” probability that a sentence belongs to a class and \( P(\text{Word} | \text{Class}) \), or the probability that a word occurs in a class. We can obtain \( P(\text{Class}) \) by calculating the proportion of sentences that belong to a class and \( P(\text{Word} | \text{Class}) \) by calculating the proportion of sentences in which a word occurs, within the class. We do not need to calculate
P(Word) because we simply take the class with the highest probability as the classification and P(Word) is the same for all classes.

Consider the example in Figure 1. Suppose that this text is the only text representing one particular class C. Now consider the word “model”. This appears twice in this excerpt of text, which has a total of 186 words. This makes P(Word=“model” | C) equal to 2/186. If the collection of texts associated with a class is small, then all such probabilities will be small. In this example the word “model” is used in different senses. There are methods that recognize different senses from the context, but in our case we just use a simple approach where space-separated objects are recognized as tokens. For instance, “New York” is also seen as two tokens: “New” and “York”. A more refined version of a classifier would recognize a compound name as a single token.

We use a version of the Naïve Bayes Classifier that uses the so-called Laplace Correction. Some probabilities of words will be zero (e.g. a word that does not occur in any sentence of a class) and this will make the product zero. We can assume that the probability of this word is in fact not zero but very small. The Laplace correction adjusts small probabilities. We use the following form of the Laplace correction:

\[
P(\text{Category} | \text{SeenWord}) = \frac{\text{Count(word)} + 1}{\text{TotalinCategory} + 2} + \frac{\text{TotalinCategory}}{\text{AllWords}}
\]

Now we can calculate the probability that a new sentence belongs to a class and assign it to the most probable class. AdaBoost uses the Naïve Bayes algorithm as a sub-procedure to construct a set of weighted classifiers. When a sentence is to be classified each classifier predicts the class and a weighted vote combines these into a single predicted class. The set is constructed by first using Naïve Bayes to build the first classifier. Next a higher weight is assigned to the sentences that were misclassified by Naïve Bayes in the first step. After that a new classifier is constructed that will, because of the weights, focus on these misclassified data. This process is repeated a fixed number (around 20) times. Each classifier that is constructed in this process receives a weight that is derived from the number of errors it makes. This combination of Boosting and Naïve Bayes gives better results than just Naïve Bayes.

With all preparatory steps (segmentation, tokenizing, etc.) taken, including the manually designed relevance classifier and the opinion classifier constructed by the machine learning algorithm, we were then ready to experiment with new reviews to find the number of sentences with a positive or negative valance for SC and WS.

5. Evaluation of the method and results

To evaluate both the machine learning method and the classifier, we used a part of the reviews to train our classifier and another part to evaluate the result. Specifically, we opted to use a procedure that is called cross validation. Our original sample of 100 AHR reviews was partitioned into three subsamples containing 33 or 34 reviews. Two of the sub-samples were used for training and the remaining ones were used as test data. This procedure is repeated three
times using each subsample once as the validation data. Overall, this gives us an indication of the accuracy of the trained classifier, or, the proportion of review sentences that the program has classified correctly. Because we use only two-thirds of our reviews to construct the classifiers the results are not as accurate as they might be if we were able to use all data for the construction process.

All steps pertaining to this method (i.e., preparation, recognition of opinion sentences, training a classifier of opinion sentences) were applied to a subset of 70 out of 100 reviews, and the result was then used to classify a different subset of 30 reviews. Classification of a review involves the following:

1. **Segmentation**: split the review into sentences.
2. **Preparation**: tokenizing, removing stop words, stemming.
3. **Relevance Classification**: classify each sentences as “opinion on SC or WS” or “neutral” (using the “opinion sentence classifier”).
4. **Opinion Classification**: classify each relevant SC and WS “opinion sentence” as positive or negative on SC and as positive or negative on WS.
5. **Calculate Review Score**: calculate the SC score and the WS score of the review from the classes of the opinion sentences.

In Table 4, the results are presented as a 'confusion matrix', which is quantified in terms of overall relevance on the basis of recall, precision and accuracy.

Table 4. Results of the relevance classifier compared to the human coder.

<table>
<thead>
<tr>
<th>Relevance Classifier: Relevant</th>
<th>Human Coder: Relevant</th>
<th>Human Coder: Non-Relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>1464</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Relevance Classifier: Non-Relevant</td>
<td>3</td>
<td>1445</td>
</tr>
</tbody>
</table>

Tables 5 and 6 present results for the sentences that were classified specifically as relevant opinion sentences. What we see from these tables is that the percentage of correct classifications (at the level of sentences) is 75% for SC and 64% for WS.

Table 5. Results of the opinion classifier compared to the human coder on sentences conveying SC+ vs. SC-.

<table>
<thead>
<tr>
<th>Scholarly Credibility (SC)</th>
<th>Human Coder: Positive</th>
<th>Human Coder: Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opinion Classifier: Positive</td>
<td>536</td>
<td>166</td>
</tr>
<tr>
<td>Opinion Classifier: Negative</td>
<td>124</td>
<td>312</td>
</tr>
</tbody>
</table>
Table 6. Results of the opinion classifier compared to the human coder on sentences conveying WS+ vs. WS-.

<table>
<thead>
<tr>
<th>Writing Style (WS)</th>
<th>Human Coder: Positive</th>
<th>Human Coder: Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opinion Classifier: Positive</td>
<td>114</td>
<td>60</td>
</tr>
<tr>
<td>Opinion Classifier: Negative</td>
<td>56</td>
<td>88</td>
</tr>
</tbody>
</table>

In the final step we used the labels of the sentences to calculate a final score of the review, based on whether it is overall positive, overall negative or overall neutral for SC and WS. The scores in Tables 7 and 8, below, were calculated using the same formulae shown in section 4.1. Here we see an accuracy of 78% for SC and 70% for WS.

Table 7. Results of program versus human on classifying the full book review as SC+ vs. SC-.

<table>
<thead>
<tr>
<th>Scholarly Credibility (SC)</th>
<th>Human coder: Positive</th>
<th>Human Coder: Negative</th>
<th>Human Coder: Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program: Positive</td>
<td>50</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Program: Negative</td>
<td>9</td>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td>Program: Neutral</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 8. Results of the program versus human on classifying the full book review as WS+ vs. WS-.

<table>
<thead>
<tr>
<th>Writing Style (WS)</th>
<th>Human Coder: Positive</th>
<th>Human Coder: Negative</th>
<th>Human Coder: Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program: Positive</td>
<td>36</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Program: Negative</td>
<td>10</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>Program: Neutral</td>
<td>5</td>
<td>2</td>
<td>24</td>
</tr>
</tbody>
</table>

6. Conclusion

We illustrate a new approach to the automated assessment of book reviews based on sentiment or opinion-related content. The results that we obtained (shown in Tables 7 and 8) are significant and reasonable for a pilot experiment; hence we are interested to see how this machine-learning approach might improve with a larger amount of training data. To obtain extensive training data more human coders would be needed to carry out the time-consuming task of coding data first by hand. This is why only 100 reviews were hand-coded for the present study. If more labeled sentences could be used as training data, then more words could be placed within our domain ontology to serve as cues for recognition. We use only two-thirds of our reviews in this experiment to construct the classifiers, thus it is unclear how much of an improvement we might see if all 100 reviews were used, or with an additional number from as much as 200 up to 1000. The method that we use relies on estimating (conditional) probabilities. If the amount of text that
belongs to a category (e.g. SC+) is small then these estimates will not be very good because many words appear infrequently. More text will improve these estimates.

Another approach to improving the results could be to split sentences that include multiple opinions into parts and label the separate parts. Note from Figure 2: “*Her penchant for generalizations will most likely irritate some readers but it is precisely this tendency that makes her book ideally suited for undergraduate courses.*” This sentence can be viewed as expressing one complex evaluation or two simpler ones. We could split this sentence here, so that the first part “*Her penchant for generalizations will most likely irritate some readers*” is negative, and the second part, “*it is precisely this tendency that makes her book ideally suited for undergraduate courses*” is positive. Although subtle meaning may be lost by splitting, this produces less ambiguous data. Without splitting, all words are associated with both a SC- and a SC+ classification, and if we split then it is clearer which words are associated with SC- and which with SC+. This would remove noisy information from the training data. Improvements might also be made to the manually constructed relevance filter. Other algorithms are often used for text classification, in particular Support Vector Machines (see for example Burges, 1998), but these are more difficult to explain and their performance is generally not much better than that of the Naïve Bayes Classifier. For large corpora where the number of relevant words is very large, Support Vector Machines may give better results.

Since our automated classifier was more accurate at recognizing positive/negative sentences related to scholarly credibility, than the same for writing style, we examined closely some of the words and phrases that reviewers used to evaluate the latter. What we noticed is that reviewers tended to use more nuanced forms of language to evaluate writing. For instance, a human coder may interpret the following statement as somewhat negative: "*[Author] writes in a pleasing narrative style that at times runs folksy.*" Within the scholarly communication system "folksy" writing might be regarded as inappropriate. On the other hand, the word "folksy" is accompanied by the phrase "pleasing narrative style", hence a sentence like this is apt to be classified (misclassified?) as positive by a computer program. It is difficult, even for a human coder, to be sure what a reviewer wants to convey with this type of comment; thus we need to observe more comments in comparison to others in a full review to obtain the overall sentiment. What does this say about writing style as an evaluative dimension? We believe that writing style is important for a history book and evidence was there to suggest that writing style and scholarly credibility are appreciated together in rhetorical assessments. In fact, there was a small correlation between the two review dimensions. Reviewers of history books may be subtler in their critique of writing style because writing style tends to be intrinsic to the person as an author. Scholarly credibility, on the other hand, can be easier to critique in a straightforward manner because it rests with external factors, such as the type of materials, methods and resources the historian has used and how thorough he or she used them to present a persuasive argument.

Earlier we suggested that the book review functions as a grand citation (i.e., ‘mega-citation’) to the quality of a newly published book/monograph. The term 'mega-citation' is a theoretical rationale for converting, or 'distilling' a qualitative piece of writing into a quantitative metric. We can count the number of reviews that a book receives as a collection of ‘mega-citations’ (i.e., one count for the review itself multiplied by the weighted sentiment) and recognize this as a measure of influence. This measure can then be used in conjunction with the regular citations that the same book receives in history journals or other books. We can also look at how often a newly
published book is purchased and catalogue in international libraries – i.e., ‘libcitations’ (White et al., 2009). Clearly a book may be reviewed or not reviewed, catalogued or not catalogued, just as it may also be cited or not cited, and ideally, a book's author would prefer that all factors are present to a high degree. Yet, because of the rhetorical richness apparent in book reviews we see value in transforming this one type of quality assessment into a weighted measure. A human agent can assist us with this calculated measure, but if an intelligent program can do it almost equally as well, a large amount of weighted megacitations could be obtained at once. Research, experimental or otherwise, will move us closer towards optimising 'mega-citations' in a comprehensive assessment of the impact of a book.

**References**


