Human Computation (2019) 6:1:147-175

ISSN: 2330-8001, DOI: 10.15346/hc.v6i1.8

Read-Agree-Predict: A Crowdsourced Approach to Discovering Relevant Primary Sources for Historians

NAI-CHING WANG, Department of Computer Science, Virginia Tech

DAVID HICKS, School of Education, Virginia Tech

PAUL QUIGLEY, Department of History, Virginia Tech

KURT LUTHER, Department of Computer Science, Virginia Tech

ABSTRACT

Historians spend significant time looking for relevant, high-quality primary sources in digitized archives and through web searches. One reason this task is time-consuming is that historians' research interests are often highly abstract and specialized. These topics are unlikely to be manually indexed and are difficult to identify with automated text analysis techniques. In this article, we investigate the potential of a new crowdsourcing model in which the historian delegates to a novice crowd the task of labeling the relevance of primary sources with respect to her unique research interests. The model employs a novel crowd workflow, Read-Agree-Predict (RAP), that allows novice crowd workers to label relevance as well as expert historians. As a useful byproduct, RAP also reveals and prioritizes crowd confusions as targeted learning opportunities. We demonstrate the value of our model with two experiments with paid crowd workers (n=170), with the future goal of extending our work to classroom students and public history interventions. We also discuss broader implications for historical research and education.

1. INTRODUCTION

Historians are often researchers as well as educators, and both roles involve significant interaction with primary sources. Primary sources are artifacts such as documents, manuscripts, diary entries, and newspaper articles created at the time under study. These sources are not only direct evidence for historical arguments (Rutner & Schonfeld, 2012) but also important materials for teaching

historical thinking skills to students in classrooms, and engaging the broader public (Stearns, Seixas, & Wineburg, 2000; Tally & Goldenberg, 2005).

As libraries, archives, and museums increasingly digitize cultural resources in their collections and share them online, historians have new and exciting opportunities to find valuable sources through web searches or even serendipitously. However, finding high-quality primary sources that are relevant to a historian's topics of interest remains a significant challenge. Because these topics are often highly abstract and specialized, they are hard to search for. For example, a scholar of the American Revolution studying "nationalism" or a Civil War historian researching "racism" and "white supremacy" would find few results searching for these keywords verbatim in transcripts, because these terms were either not used at the time or require a modern perspective to apply them to historical documents. Manual indexing by professional metadata librarians is cost-prohibitive for many organizations, but when it is available, annotations are often provided to serve the broadest possible audience. Automated approaches to text analysis also struggle to provide relevant results for these "long tail" searches with long semantic distances from the source material. Consequently, historians often end up devoting large amounts of time to manually evaluating the relevance of the contents of these archives. Not surprisingly, they are often frustrated at spending so much time on tasks other than writing and analysis (Rutner & Schonfeld, 2012).

Crowdsourcing could provide an alternative approach to overcoming these challenges. Crowdsourcing has been shown to be effective for many types of text analysis, from transcription (Little, Chilton, Goldman, & Miller, 2010) to word processing (Bernstein et al., 2010) to qualitative analysis and clustering (André, Kittur, & Dow, 2014). However, little research has sought to use crowdsourcing to perform in-depth analysis of historical documents for the purposes of labeling relevance. We suggest one key problem is that novice workers employed on popular crowdsourcing platforms like Amazon Mechanical Turk (MTurk) typically lack the expertise in history that is presumably necessary for such judgements.

In this article, we present a crowdsourcing approach that enables novice crowds to label the relevance of digitized primary sources as accurately as expert historians. To develop this approach, we first conducted a preliminary experiment with 120 MTurk crowd workers and a real-world online archive of digitized American Civil War-era documents, in collaboration with professional historians. This study investigated which of three interface designs, based on theories from educational psychology, would best support crowdsourced relevance labels. Informed by these results, we developed our crowdsourcing approach, which we call Read-Agree-Predict (RAP).

With RAP, a historian first provides her specialized topic of interest and two example documents. Then, crowd workers use a novel interface to read historical documents and label their relevance to the historian's topic. Finally, the results are aggregated based on worker agreement to produce a final relevance label to the historian for each document. Our analysis of the preliminary study data found that RAP enabled perfect precision and recall for labeling relevance, outperforming both a

majority vote aggregation and individual worker performance. RAP is simple enough to be easily adapted for most online archives. As a useful byproduct, RAP identifies areas of crowd confusion that could help historians prioritize teaching opportunities in classrooms or public history projects.

To validate RAP, we conducted a second experiment with 50 additional MTurk workers, new historical documents, a new topic, and a new expert historian. We again found that RAP enabled the novice crowds to label relevance as well as an expert. Additional validations include a simulation study of RAP with different crowd sizes, a comparison of relevance agreement between two historians, and a comparison to automated text analysis approaches for labeling relevance.

Our contributions in this article include the technical contribution of the RAP crowdsourcing approach and the empirical contributions of preliminary and validation studies demonstrating the effectiveness of RAP on real-world digitized historical documents. As a result, historians can spend more time analyzing and interpreting primary sources, rather than searching for them, and consider to a larger set of relevant documents than they would have been able to locate on their own. The results may have implications for historical scholarship and history education.

2. RELATED WORK

2.1 Challenges in Historical Research

Studies of the practices of historians and research support professionals show that interacting with primary sources, including gathering, discovering, and organizing historical documents, remains central to historical research (Dalton & Charnigo, 2004; Nawrotzki, 2013; Wineburg, 2010). When gathering sources, historians identify and assess relevant sources to address their research questions and support arguments, and organize the sources primarily based on their own specialized topics of interest. Later, they can use these topics to find the sources again (Rutner & Schonfeld, 2012). Historians spend a large part of their daily work gathering and discovering sources relevant to their specialized research topics (Dalton & Charnigo, 2004; Rutner & Schonfeld, 2012).

Even with modern search engines, historians often cannot directly search by topics of interest (Rutner & Schonfeld, 2012). This is because these topics are often not keywords found verbatim in the raw texts, and a set of keywords might have different relevance to topics depending on the context. Existing archives may have rich annotations created by metadata librarians or other professionals, but because these require time and expertise, they tend to focus on topics of broader interest to maximize their utility.

Historians may try to overcome these limitations by using many different search terms to find resources relevant to their unique topics of interest, and filter out many irrelevant search results caused by wrong search terms. Historians may conduct individualized foraging because the topics

historians look for and the approaches historians take may be very different from one another. This individualized foraging and curation process is time-consuming and sometimes tedious, but seen as an inescapable part of the historical research process. Some historians even regret spending the time on organizing instead of other research activities. As one historian interviewee said, "*Once it's organized, it's up to me to think about it and write. But I do resent the time that's spent organizing and managing everything*" (Rutner & Schonfeld, 2012). In this article, we consider how novice crowds could support individual historians to improve the breadth and efficiency of their organizing efforts.

2.2 Crowdsourcing in the Classroom

Crowdsourcing research has started to explore the use of crowdsourcing in classroom-related settings. These studies often aim to address the issue of low ratios of instructors to students, especially in massive open online courses (MOOCs), by leveraging peer learners or other (sometimes paid) crowds to provide feedback and improve learning through collaboration. Some of this work seeks to enhance learning with collective learner activity. For example, while watching a teaching video, students may pause at different places to digest the content. The aggregated pause positions may hint at important or confusing parts of the video. Other work seeks to structure the learning process or provide learners with useful feedback from other (e.g., paid) crowds. These efforts include creating crowdsourced sub-goals in how-to videos (Kim, Miller, & Gajos, 2013; Kim & others, 2015; Weir, Kim, Gajos, & Miller, 2015), crowdsourced assessments or exercises (Mitros, 2015; Šimko, Šimko, Bieliková, Ševcech, & Burger, 2013), personalized hints for problem-solving (Glassman, Lin, Cai, & Miller, 2016), receiving design critiques (Xu, Rao, Dow, & Bailey, 2015), and identifying students' confusions (Glassman, Kim, Monroy-Hernández, & Morris, 2015).

Studies of historians' current practices in classrooms show that while historians may have very different ways of teaching history, such as chronological narration or topic-based class activities, lecturing continues to be the established practice (Grant, 2001, 2018; Grant & Gradwell, 2010; McDaniel, 2010). Historians rarely include crowdsourcing to integrate classroom teaching with their research. Other research has found that scholars may resist crowdsourcing for research due to knowledge and role uncertainties (Law, Gajos, Wiggins, Gray, & Williams, 2017). For example, one historian expressed a need to know the person and have some human link in order to trust the quality of the data they produce. In addition, there may be a moral dilemma in asking students to do unattractive tasks which could be perceived as exploitative without promoting learning (Law, Gajos, Wiggins, Gray, & Williams, 2017). These concerns are consistent with issued raised in discussions with our historian collaborators.

Like some of this prior work, we seek to leverage crowdsourcing to support classroom learning, but we extend this thread of research in several ways. First, we explore this goal in the domain of history, which has seen little attention among crowdsourcing researchers to date. Second, our approach is designed to work in a research context, where the answers are not known a priori and

the crowd contributes authentically to the historian's scholarship. While our long-term vision is that historians could deploy our approach with students in their classrooms, this article focuses on a proof-of-concept with paid crowd workers. The results of our study may help mitigate the uncertainties some historians have about adopting this technique in the classroom.

2.3 Crowdsourced Text Classification and Labeling

Many automated techniques have been developed to handle the task of both single-label and multi-label text classifications (Aggarwal & Zhai, 2012; Sebastiani, 2002; M.-L. Zhang & Zhou, 2014). However, recent studies show that state-of-the-art automated techniques may still be far from perfect (Venkatesan, Er, Dave, Pratama, & Wu, 2016; X. Zhang, Zhao, & LeCun, 2015). In addition, in order to perform well, these techniques often require many examples with high-quality labels as training data (e.g. (Banko & Brill, 2001; Kavzoglu & Colkesen, 2012)). Therefore, a large body of research focuses on how to generate high-quality labels, e.g., using experts or crowdsourcing.

While experts can produce high quality labels, experts are often rare and expensive, as in the case of historical research. On the other hand, while crowds can produce a much larger number of labels, they often lack required domain expertise. Therefore, much crowdsourcing research explores and develops aggregation techniques to increase the quality of crowdsourced labels from non-expert crowds in various tasks, such as affective text analysis, word similarity, and word sense disambiguation (Snow, O'Connor, Jurafsky, & Ng, 2008). Majority vote aggregation is shown to be effective when the crowd's responses are imperfect but better than chance (Sheng, Provost, & Ipeirotis, 2008; Snow et al., 2008). With work history, expectation maximization (EM) (Dawid & Skene, 1979) can generally be used to further improve quality (Hosseini, Cox, Milić-Frayling, Kazai, & Vinay, 2012; Ipeirotis, Provost, & Wang, 2010; McDonnell, Lease, Elsayad, & Kutlu, 2016; Snow et al., 2008), although it may only converge on local maxima instead of achieving global optima (Drapeau, Chilton, Bragg, & Weld, 2016). With work history, it is also possible to improve the quality of results by considering individual workers' systematic biases (Giancola, Paffenroth, & Whitehill, 2018). For example, using only majority vote, accuracy ranges from 0.58 to 0.80, whereas using EM, the accuracy increases to a range of 0.64 to 0.82 for the INEX dataset (Hosseini et al., 2012). Identifying workers with domain expertise can also provide better quality results (Drapeau et al., 2016; Prelec, Seung, & McCoy, 2017). Studies also show that appropriate hierarchical schemes and task assignments can also improve quality of the results for multiclass classification (Duan & Tajima, 2019).

We contribute to this literature by exploring how crowdsourced labeling and classification can be used to contribute directly to historical research. Unlike many automated approaches, our model requires minimal training data (two example documents) and works as well as experts for abstract, long-tail search topics. Different from the aforementioned crowdsourcing research, which requires work history or special ways of identifying expertise, our model works well with typical paid crowds that lack specialized knowledge and have short time commitments. However,

crowds with more training and motivation (e.g., students or enthusiasts) might perform more efficiently.

2.4 Educational Psychology and Semantic Tasks

Making connections between topics and relevant documents in order to label relevance requires readers to have a good understanding of both. Reading comprehension has been modeled as a complex cognitive process involving different levels of lexical and semantic processing (Kintsch & van Dijk, 1978). Research on levels of processing suggests that deeper elaboration leads to better recall and understanding (e.g. (Craik & Lockhart, 1972; Craik & Tulving, 1975)). Underlining and summarizing are semantic tasks that may trigger deeper levels of processing with more elaboration and thus increase reading comprehension (Bobrow & Bower, 1969; Doctorow, C, & Marks, 1978; Linden & Wittrock, 1981; Schnell & Rocchio, 1978; Smart & Bruning, 1973; M. C. Wittrock & Alesandrini, 1990). Improved reading comprehension may help novices apply better relevance labels. In our preliminary study, we explore how different types of semantic tasks can affect task performance in a crowdsourced history context.

3. PRELIMINARY STUDY

3.1 Research Questions and Hypotheses

Drawing on the above literature review, we sought to understand how to design an interface for crowds to effectively label the relevance of historical primary sources to topics of interest. Therefore, we conducted a preliminary experiment to establish a quality baseline for crowdsourced contributions in the domain of history. This study compared two popular reading comprehension techniques, underlining and summarizing, with a reading-only (baseline) interface. We refer to these three techniques as semantic tasks. We hypothesize that summarizing will have a stronger effect on performance than underlining or reading because summarizing has been shown to require the deepest level of processing in writing tasks (Cai, Iqbal, & Teevan, 2016). Specifically, our preliminary study explored the following research questions and hypotheses:

RQ1: How does the semantic task (reading, underlining, or summarizing) affect the quality of crowd-generated relevance labels?

H1: The quality of crowd labels will be highest in the summary task and lowest in the reading task.

RQ2: How does the semantic task affect the agreement of crowd-generated relevance labels?

H2: The agreement of crowd labels will be highest in the summary task and lowest in the reading task.

RQ3: How does the semantic task affect the efficiency of applying crowd-generated relevance labels?

H3: The efficiency of applying crowd labels will be lowest in the summary task and highest in the reading task.

3.2 Dataset and Historian

The documents used in this study come from a digital archive[1] of 189 digitized historical primary sources (personal diaries and letters, newspaper articles, and public speeches) from the American Civil War era (ca. 1840-1870). This archive was assembled by a tenured professor of Civil War history at our institution (Historian A, also the third co-author of this article) for a prior research project. Historian A generated a list of six topics of interest, related to Independence Day celebrations, that he used to build the archive. We used a subset of these documents and topics for this study, as detailed in the Experimental Design (Section 3.5).

3.3 Apparatus and Procedure

The experiment was conducted entirely online. After completing an online IRB-approved consent form, each participant (an MTurk worker) was randomly assigned to one of three conditions corresponding to one of the three semantic tasks: reading, keyword (underlining), or summary (summarizing). While prior studies on underlining were often conducted in pen-and-paper settings, our study takes place online, so we instead asked participants in the keyword (underlining) condition to type in a set of keywords. While both activities similarly ask participants to identify and highlight important words and phrases, selecting keywords is a more common (and arguably, more natural and familiar) task on MTurk than underlining. Each participant was also assigned a topic and a document. The participant then used the web interface we developed, based on a few alternative designs in pilots, to complete a three-step process as shown in Figure 1.

First, the participant filled out a short quiz in which they matched their assigned topic to its correct definition. If participants did not get the answer right, they could not proceed to the next step, and had to end the task themselves. This step ensured all participants in the study understood the topic's meaning. We did not observe any cases where the participant proceeded without providing the correct answer.

Next, the participant viewed two example documents for their topic with relevance labels provided by Historian A. Our pilots and recent work on crowd innovation (Yu, Kittur, & Kraut, 2014) both suggest that by viewing good examples, people can better understand abstract concepts and analogies. The participant also practiced their assigned semantic task on these examples. The reading task involved simply reading the example documents. The keyword task involved reading the documents and selecting 4-8 important keywords or phrases for both. The summary task involved reading the documents and writing a 1-2 sentence summary for both.

[1] http://www.july4.civilwar.vt.edu

Third, the participant completed the semantic task on a new document. After completing the task, the participant decided whether it was relevant to the assigned topic by clicking "Yes" or "No" and typing in a brief justification of their decision.

3.4 **Participants**

We used Amazon Mechanical Turk to recruit novice crowd workers. We restricted to US-only workers to increase the likelihood of English language fluency, with a 95% HIT (human intelligence task) minimum acceptance rate and 50 or more completed HITs. We recruited 120 workers and randomly assigned 40 to each of the three conditions. Each worker was unique and assigned to only one HIT to ensure that the required expertise was learned within that HIT. Thus, there were five unique workers for each combination of condition (semantic task), document, and topic. We paid participants $7.25/hour based on average task times in pilots. We also paid them a 20% bonus payment if they provided a reasonable justification for their decision, even if it was wrong. The total amount paid to workers for the preliminary study was about $192, including $32 in worker bonuses but excluding MTurk platform fees.

Although learning is often assessed with students in classrooms, a risk is that new teaching methods may hinder students' learning if the methods are not effective (Brown, 1992). Our studies use paid crowds on MTurk in order to validate our approach in a controlled lab setting where participants are compensated regardless of how much they learned. In Section 6.2.1, we discuss how our findings from MTurk studies could be adapted for students in the classroom.

3.5 **Experimental Design**

This was a between-subjects design with one independent variable (semantic task), two covariates (topic and document), and three dependent variables (quality, agreement, and efficiency).

3.5.1 Independent Variable

The independent variable, semantic task type, had three levels: reading, keyword, or summary. Therefore, the experiment had three conditions.

Figure 1: User interface for the summary condition in the preliminary study.

3.5.2 Covariates

We controlled for two covariates: topic and document. The complexity of the topic is likely to affect crowd performance, so we selected four diverse topics — *Revolutionary History and Ideals*, *American Nationalism*, *American Hypocrisy*, and *Anxiety* — from the list generated by Historian A. Table 1 shows the number of relevant documents for each of the four topics across the entire dataset of 189 documents. Each document was labeled for all four topics; a document may be relevant to multiple topics.

Document complexity could also affect crowd performance. Therefore, we randomly selected documents that were similar in terms of length (mean=265 words) and readability (college-level, according to Flesch-Kincaid readability tests). We randomly selected two documents for each topic, one highly relevant and one irrelevant (as judged by Historian A), for a total of eight documents. None of the documents contained the topic name verbatim.

Table 1: Number of relevant documents for each topic according to Historian A.

Historical Topic	**Relevant Documents**
Revolutionary History and Ideals	21
American Nationalism	38
American Hypocrisy	10
Anxiety	28

3.5.3 Dependent Variables

There were three dependent variables: quality, agreement and efficiency. To measure *quality*, we compared how each worker's responses compared with gold standard responses provided prior to the study by Historian A. Specifically, we measured the accuracy, precision, and recall of the labels applied by the crowd, i.e., whether they indicated a document was relevant or irrelevant to their assigned topic. These metrics are widely used in the field of Information Retrieval to measure the performance of an information retrieval system. We measured accuracy as the ratio of correct labels (both relevant and irrelevant) to the total number of labels applied by the crowd. This gave us an overall idea how close the crowd's labels were to the experts' by considering both relevant and irrelevant labels. We measured precision as the ratio of the number of correct relevant labels to the total relevant labels indicated by the crowd. Precision can be seen as an indicator of how credible the crowd's labels were. We measured recall as the ratio of the number of correct relevant labels to Historian A's relevant labels. This measure told us if the crowd missed any relevant documents.

We also measured *agreement* among the five workers assigned to each condition. This metric provides an indicator of reliability for crowd workers and identifies areas of confusion as

potential teaching opportunities. We used two measures of agreement, Fleiss' κ (Fleiss, Levin, & Paik, 2013) and Raw Agreement Indices (RAI) (Fleiss et al., 2013; John Uebersax, 2009). Fleiss' κ provides overall agreement and there is some established interpretation for its values. In addition to overall agreement, RAI also allows finer-grained calculations, such as an agreement value for a particular document in a condition. Both Fleiss' κ and RAI use a 0-1 scale where 0 is no agreement and 1 is perfect agreement.

We also measured the crowd's *efficiency* in analyzing documents in terms of both time and attempts. Time describes how long it takes for a task to be completed and is a measure of how much effort the task requires. Attempts (attrition) describes how many workers accept and return a HIT before it is completed and is a measure of the perceived difficulty of the task.

3.6 **Results**

3.6.1 Individual Quality Similar Across Conditions

There was no significant difference in individual quality across the three conditions in terms of accuracy, precision, or recall. The results of the quality analysis are shown in Table . The average accuracy across all conditions was 0.68 (max: 1.0). The average accuracy per condition was reading: 0.70, keyword: 0.68, and summary: 0.65. A one-way ANOVA showed semantic activity did not have a significant effect on accuracy ($F(2, 21)=0.051$, p=n.s.). The average precision values for the reading (0.64), keyword (0.64), and summary (0.62) conditions were very similar. The average recall values were 0.90 for reading and 0.80 for both keyword and summary.

3.6.2 Majority Vote Improves Quality

Since we had five unique worker results for each combination of condition, document, and topic, we also considered how an aggregated (majority vote) decision affected quality. When we used a majority vote strategy, there was only one false positive for the keyword condition and two false positives for each of the other two conditions, giving overall accuracy values of 0.88 and 0.75, respectively. The precision values are 0.80 for keyword and 0.67 for both reading and summary. The recall value is 1.0 for all three conditions.

3.6.3 Summarizing Leads to Higher Agreement

We found that the summary condition led to higher average agreement. Table shows intra-crowd agreement. Both measures, Fleiss' κ and RAI, showed very similar results and trends (with Pearson correlation coefficient r=1.0). For RAI, average agreement was highest in the summary condition (mean=0.83). Agreement was similar in the reading (mean=0.60) and keywords (mean=0.58) conditions. For agreement in individual documents, a one-way ANOVA showed no effect of semantic task on RAI scores ($F(2, 21)=2.676$, p=n.s.).

For Fleiss' κ, average agreement in the summary condition was 0.80, interpreted as between "substantial agreement" and "almost perfect agreement." The κ values for the reading and keyword conditions were similar, 0.56 and 0.54 respectively, indicating "moderate agreement."

*Table 2: Quality results for the preliminary study. There were 5 workers per doc and each worker labeled only 1 doc). * indicates the average refers to accuracy only.*

		Topic 1		**Topic 2**		**Topic 3**		**Topic 4**		**Avg.**
Document	**Condition**	1	2	3	4	5	6	7	8	
Relevant? (Historian A)		N	Y	N	Y	N	Y	N	Y	
Relevant? (Majority Vote)	Reading	Y	Y	Y	Y	N	Y	N	Y	0.75*
	Keyword	Y	Y	N	Y	N	Y	N	Y	0.88*
	Summary	Y	Y	Y	Y	N	Y	N	Y	0.75*
Relevant? (RAP)		N	Y	N	Y	N	Y	N	Y	
Crowd Labels	Reading	YYYNN	YYYYY	YYYNN	YYYYN	YYNNN	YYYYY	YYNNN	YYYYN	
	Keyword	YYYYN	YYYYN	YYNNN	YYYYY	YNNNN	YYYYN	YYNNN	YYYNN	
	Summary	YYYYY	YYYYN	YYYYY	YYYNN	NNNNN	YYYYY	NNNNN	YYYYN	
Crowd Accuracy	Reading	0.4	1.0	0.4	0.8	0.6	1.0	0.6	0.8	0.70
	Keyword	0.2	0.8	0.6	1.0	0.8	0.8	0.6	0.6	0.68
	Summary	0.0	0.8	0.0	0.6	1.0	1.0	1.0	0.8	0.65
Crowd Precision	Reading									0.64
	Keyword									0.64
	Summary									0.62
Crowd Recall	Reading									0.90
	Keyword									0.80
	Summary									0.80

*Table 3: Intra-crowd agreement results for the preliminary study. * indicates teaching opportunity.*

		Topic 1		**Topic 2**		**Topic 3**		**Topic 4**		**Avg.**
Document	**Condition**	1	2	3	4	5	6	7	8	
Raw Agreement Indices (RAI)	Reading	0.4	1.0	0.4	0.6	0.4	1.0	0.4	0.6	0.60
	Keyword	0.6	0.6	0.4	1.0	0.6	0.6	0.4	0.4	0.58
	Summary	1.0*	0.6	1.0*	0.4	1.0	1.0	1.0	0.6	0.83
Fleiss' κ	Reading									0.56
	Keyword									0.54
	Summary									0.80

3.6.4 Reading is fastest

Overall, the average time to complete a task was about 11 minutes (SD = 5.3, min=1.5, max=29). Broken down by condition, the averages were reading: 7.8 min (SD = 3.7), keyword: 11 min (SD = 4.5), and summary: 13 min (SD = 6.0). A one-way ANOVA showed condition had a significant effect on time ($F(2, 117)=12.66$, $p<0.01$). Post-hoc Tukey tests showed that the reading condition was significantly faster than both the keyword and summary conditions. There was no difference between the keyword and summary conditions.

3.6.5 Keywords Require Most Attempts

Overall, on average, it required about 2.8 attempts (SD = 2.2, min=1, max=11) to complete a task. Average attempts per condition were reading: 2.15 (SD = 1.7), keywords: 3.80 (SD = 2.7), and summary: 2.55 (SD = 1.9). A one-way ANOVA showed that condition had a significant effect on attempts ($F(2, 117)=6.65$, $p<0.01$). Post-hoc Tukey tests showed that it took significantly more attempts to complete the keyword condition than the reading and summary condition. There was no difference between the reading and summary conditions.

3.7 **Discussion**

3.7.1 Quality

The results of the quality analysis showed that semantic task did not affect quality for individual workers, so H1 is refuted. Across all three conditions, individuals in the crowd did better than flipping a coin but might not perform well enough for scholarly work. This result supported the general assumption that a novice might not be able to produce high quality results due to a lack of expertise. The quality scores indicate occurrences of crowd confusions, while the reasoning provided by the crowd elaborates on what the confusions were.

By investigating the keywords in the keyword condition, we found that some participants applied wrong labels based on just a few keywords. For example, when some participants saw the keyword "Fourth of July", they directly connected the document to the topic *Revolutionary History and Ideals* regardless of the context for how "Fourth of July" was used. This outcome may be evidence of the von Restorff effect, i.e., when multiple similar stimuli are present, the one that differs from the rest is most likely to be remembered. Previous studies have identified this phenomenon in underlining or highlighting because participants tend to remember what has been highlighted (E. H. Chi, Hong, Heiser, & Card, 2006; Ed H. Chi, Hong, Gumbrecht, & Card, 2005; Nist & Hogrebe, 1987; Peterson, 1991).

However, measuring aggregated crowd results using a *majority vote* technique showed better results, in line with prior work (e.g., (Sheng et al., 2008)). The higher recall and precision values suggest real-world potential for crowds supporting historians because workers were able to find all relevant sources while filtering out some irrelevant ones. For example, in the preliminary

study dataset, crowds could have reduced the size of search pool for the historian by 37.5% in the keyword condition (correctly eliminating 3/4 irrelevant sources, leaving 1 false positive) or 25% in the reading or summary conditions (each correctly eliminating 2/4 irrelevant sources, leaving 2 false positives). Thus, 75% of time the historian spent reviewing irrelevant documents could have been saved in the keyword condition, and 50% of time the historian spent reviewing irrelevant documents could have been saved in the reading and summary conditions.

3.7.2 Agreement

The results for intra-crowd agreement shown in Table partially support H2. While there was no significant difference among the conditions based on the fine-grained RAI agreement value for each document, the overall agreement was higher at the summary condition (0.83) than in the reading (0.60) and keyword conditions (0.58). This result seems to be in line with previous studies (Bretzing & Kulhavy, 1979; Cai et al., 2016; Doctorow et al., 1978; M. C. Wittrock & Alesandrini, 1990; Merlin C. Wittrock, 1989) showing that summarizing demands a deep level of semantic processing.

When multiple crowd workers make the same incorrect labels, it often means there is some shared confusion or common misunderstanding. This situation suggests an opportunity for historians to help the crowd better understand the material. Like most experts, historians' time is limited, so it is important to prioritize these misconceptions to help as many workers as possible. Our measure of intra-crowd agreement can be a good indicator for this.

The results from Table showed that there were two high-impact confusions in the summary condition for Topic 1 and Topic 2. In these situations, the crowd majority thought an irrelevant document was relevant. For example, for Topic 1 (*Revolutionary History and Ideals*), Historian A did not think it was relevant, but crowd workers in the summary condition all thought it was. The example is shown in Table along with two of the crowd workers' responses and Historian A's feedback.

Table 4: Sample crowd misconceptions for Revolutionary History and Ideals *in the summary condition.*

Unrelated Testing Document (The National Anniversary, *The Daily Dispatch*, 07/03/1860)	**Participants' Reasons (Summary Condition)**
"There seems to be preparations going on in all the principal cities of the Union to celebrate the Fourth of July in the old-fashioned style of military, oratorical and patriotic jubilation. There is a good deal of American feeling still left in the country, and it makes itself manifest on all suitable	"It is related to the topic of *Revolutionary History and Ideals* through the language used within the document (words, such as Union, sectionalism and political parties), the mention of old fashioned military, and the general timeless sense of patriotism and national pride." – P21

<table>
<tr><td rowspan="2">occasions. It is pleasing to observe that all the political parties emphatically announce their loyalty to the Union, which is a strong proof that sectionalism is not popular. Far distant be the day when the Fourth of July shall awaken no patriotic associations, sentiments and hopes in the breasts of American citizens!”</td><td>Historian A’s feedback: Sectionalism does not fit the topic as well as some of the other words.</td></tr>
<tr><td>“It speaks to the unity that Americans feel. It's a matter of pride. And it's always going to be. July 4th is always going to be central in the hearts of all Americans. It's a day to celebrate because this country has done so much good for so many people” – P24

Historian A’s feedback: This knee-jerk reaction to the July 4 reference introduces emotions not present in the document.</td></tr>
</table>

3.7.3 Efficiency

The results partly supported H3 in that the reading condition was significantly faster than the other two. However, with respect to number of attempts, there was no difference between reading and summary, while keywords required significantly more attempts than the others. This latter result surprised us, as summarizing has been previously shown to be more cognitively demanding. One possible explanation is that our instructions were phrased in a way that made the keyword condition seem more laborious than it actually was. In the keyword condition, participants were asked to provide “4-8 keywords/keyphrases” while in summary condition, participants were asked to provide “1-2 sentences.” By glancing the numbers shown in the semantic task instructions, there may have seemed more work to be done in the keyword condition than for the summary condition.

4. READ-AGREE-PREDICT (RAP)

The preliminary study showed mixed results for the three semantic tasks: reading, underlining, and summarizing. Going beyond the original research questions, we made several observations in our follow-up data analysis that suggested an approach that could yield better results than any one task, and better than other common aggregation techniques like majority vote. We call this combined approach Read-Agree-Predict (RAP).

4.1 Observations from Preliminary Study

In the preliminary study, there were three possible levels of intra-crowd agreement: zero workers vs. five, one vs. four, and two vs. three, corresponding to RAI scores of 1.0, 0.6, and 0.4, respectively (see Table). While the first two levels were considered high agreement because the crowd had a clear majority choice, the third was considered low agreement because workers were nearly equally split. We could therefore choose 0.6 as a threshold to distinguish high (≥0.6) and low (<0.6) agreement.

We made two observations with respect to this agreement threshold that held for only the *reading condition*. First, we observed that if crowd agreement was low (RAI=0.4), the document was always irrelevant. In other words, confusion or disagreement among workers suggests the document is not relevant to the topic. These situations may reflect ambiguity or a lack of information in the source material.

Second, we observed that if crowd agreement was high (RAI≥0.6), the crowd's majority-vote decision was highly accurate. In other words, when crowds converge on a single decision (relevant or irrelevant), that decision could usually be trusted. These situations may occur when there is sufficient evidence for the crowd to make a clear yes-or-no decision.

Taken together, these observations about the reading condition suggest the following robust pattern could be used to be used to predict highly accurate labels between documents and topics. If a crowd reading a document reaches low agreement about its relevance to a given topic, i.e., a nearly split vote over whether the document is or is not relevant, then we can predict the document is irrelevant. However, if a crowd has high agreement about a document's relevance, its majority vote decision (relevant or irrelevant) can be trusted. We call this pattern Read-Agree-Predict (RAP).

4.2 RAP vs. Majority Vote

RAP can be viewed as an improvement upon majority vote for crowdsourced adjudication. This improvement is two-fold. One, it tells when to reliably use majority vote — only when crowd agreement is high (RAI≥0.6). Two, it offers a judgement when majority vote is not reliable — the document is irrelevant to the topic. While much crowdsourcing research uses simple majority vote for adjudication or relevance assessment, RAP pushes the concept a step further by 1) demonstrating how a threshold value of majority may have strong impact on output, and 2) providing a clear binary relevance judgement in all possible situations.

Overall, in the preliminary study, majority vote allowed the crowd to achieve quality scores up to 0.8 (precision) and 1.0 (recall) for certain topics and documents. These results could have helped reduce the size of a historian's search pool by up to 37% and saved up to 75% of time spent on irrelevant documents in the archive.

For comparison, we applied RAP post-hoc to the preliminary study's dataset. The results in Table show that RAP is a substantial improvement over majority vote, yielding perfect accuracy relative to Historian A's gold standard judgements. RAP achieved scores of 1.0 (precision) and 1.0 (recall) across all documents and topics. These results suggest a historian would not even have to search the digital archive herself, because crowd workers using RAP would have correctly labeled all relevant documents.

4.3 Crowd Confusion as Teaching Opportunities

Beyond producing high quality labels from noisy ones, RAP not only detects where students' confusions may occur but also prioritizes these confusions as a useful byproduct. If agreement in the reading condition is low and the majority thinks the unknown document is relevant, then RAP predicts this source-topic pair will be a high-impact confusion for teaching. We further discuss educational opportunities in Section 6.2, with a usage scenario in 6.2.1.

In the next section, we simulate this usage scenario with a new study to validate RAP.

5. VALIDATION STUDY

We conducted this study to validate RAP, so the experimental design was almost identical to the preliminary study. We summarize the differences below.

5.1 Methods

5.1.1 Dataset and Historian

We again used the same online archive of American Civil War primary sources as in the preliminary study, but it had since been expanded to about 1200 primary sources. For the validation study, we randomly sampled 10 new documents of similar length and readability level that were not among the set of eight used in the preliminary study. Next, we recruited a new expert historian from our institution, Historian B. We asked Historian B, without seeing the 10 documents, to provide a topic of interest, a definition, and two historical documents that were good examples of that topic. Drawing on his research interests, he chose the topic "Racial Equality." Finally, we asked Historian B to generate gold standard answers by reading each of the 10 documents and deciding whether it was relevant or irrelevant to his topic.

We used this selection mechanism because 1) it avoided biasing our expert, and 2) it reflected how RAP would be used in a real-world situation. That is, a historian locates an unfamiliar digital archive, provides a topic, definition, and two example documents, and the crowd analyzes each document from the archive to decide if it is relevant to that topic. After that, the historian comes back to check the sources labeled by the crowd.

5.1.2 Apparatus and Procedure

We used a very similar web-based interface and procedure as the preliminary study. We kept the reading condition the same as before. We removed the keyword and summary conditions, which were significantly slower than reading yet comparable in terms of quality.

5.1.3 Participants

We recruited 50 participants on Amazon Mechanical Turk using the same criteria and pay rate as the preliminary study. The total amount paid to workers for the validation study was about $56, including $19 in worker bonuses but excluding MTurk platform fees.

5.1.4 Experimental Design

The experimental design mirrors that of the preliminary study, with the exception of document selection procedure described in Section 5.1.1.

5.2 **Results and Discussion**

After collecting the crowd data from 50 workers, we ran the data through the RAP crowd algorithm as well as a standard majority vote aggregation to generate predictions of relevance for each of the 10 documents. Table shows that majority vote yielded perfect recall but 2 false positives, similar to the majority vote performance for the reading condition in the preliminary study. In contrast, the RAP predictions exactly matched the gold standard answers provided by Historian B. Thus, in this validation study, RAP again achieved perfect accuracy for a new historian, new topic of interest, and new random sample of documents within the same digital archive as in the preliminary study. RAP also automatically prioritized documents with crowd confusions based on the number of wrong votes for the historian's reference.

We also noticed that the ratio of relevant documents in the validation study was much higher than that of the preliminary study. We speculate this is partly due to the topic chosen by Historian B. Since Historian B's topic was "Racial Equality," and African-American slavery was a main cause of the American Civil War, this topic may have a higher relevance ratio than more specialized topics, such as those used in the preliminary study.

*Table 5: Quality and agreement results for the validation study. * indicates teaching opportunity.*

		Topic 5									
	Document	9	10	11	12	13	14	15	16	17	18
Quality	**Relevant? (Historian B)**	N	N	N	Y	N	Y	Y	Y	N	N
	Crowd Labels	YYYNN	YYNNN	YNNNN	YYYYY	YYYNN	YYYYY	YYYYN	YYYYY	NNNNN	YYNNN
	Relevant? (Majority Vote)	Y	N	N	Y	Y	Y	Y	Y	N	N
	Relevant? (RAP)	N	N	N	Y	N	Y	Y	Y	N	N
Agreement	**RAI**	0.4*	0.4*	0.6	1.0	0.4*	1.0	0.6	1.0	1.0	0.4*
	Wrong Votes (out of 5)	3	2	1	0	3	0	1	0	0	2

5.3 Simulating Different Crowd Sizes

To further investigate the effectiveness of RAP, we ran a simulation to understand how RAP would compare to majority vote with different hypothetical crowd sizes. For each crowd size *n*, we resampled (with replacement) the existing crowd data to create the desired crowd size. We then calculated the average F-1 score for 1,000 resampled data points. We used the F-1 score (harmonic mean of precision and recall) because it is a widely-used measure of search performance in Information Retrieval research.

In Figure 2, "Ideal Average F-1 Score" is the best average F-1 score that RAP achieves for a given crowd size. "Cut-off Agreement for Ideal Average F-1 Score" is the recommended agreement threshold to achieve the ideal average F-1 score. "Average F-1 Score with Cut-off Agreement = 0.6" is the average F-1 score using a threshold of 0.6. "Majority Vote (Reading)" is an F-1 score using the majority vote from the reading condition. Crowd size is on the x-axis, and both agreement threshold and F-1 score are shown on the y-axis.

The simulation results suggest three key takeaways. First, RAP's average F-1 score is very close to the ideal average F-1 score (correlation coefficient=0.99). This suggests that the agreement threshold we used for both the preliminary and validation studies, 0.6, was an effective choice.

Second, RAP with either 0.6 or the ideal agreement threshold outperforms simple majority vote for all crowd sizes. The highest majority vote score is still worse than the lowest possible RAP score, which occurs at crowd size=3.

Third, the benefits of RAP increase with larger crowd sizes, approaching perfect accuracy. At crowd size=5, used in the preliminary study, the average performance of RAP is already close to Historian B, with F-1=0.84. At crowd size=11, the average performance is equal to Historian B, with F-1=0.89. In contrast, the F-1 score for majority vote quickly saturates at around 0.8.

Finally, we note that if the worker accuracies change a lot, the threshold will also change, but the process to find the threshold would be similar. We would also expect the threshold to be slightly different in other domains, depending on how similar those domains are to history.

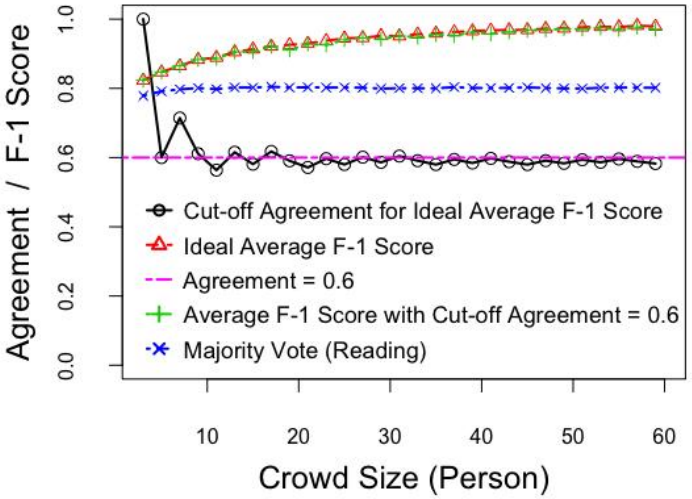

Figure 2: Comparison of agreement methods and recommended agreement vs. crowd size in validation.

5.4 Historian Accuracy and Agreement

To complement this validation, we also sought to create a baseline of historian performance by comparing the two historians in our studies, Historian A and Historian B, to each other. We asked both historians to judge relevance for the document set they had not seen before. Historian B judged Documents 1-8 (preliminary study), and Historian A judged Documents 9-18 (validation study). Across both document sets, there was substantial agreement between the two historians (Cohen's $\kappa = 0.72$). The average F-1 score across both historians and document sets was 0.89. This could be interpreted as a general measure of historians' performance in finding relevant sources for other historians.

These results support the intuition that historians can have slightly different interpretation of documents based on their research context. RAP is able to follow individual historians' interpretation in their individual research context by achieving perfect accuracy for both historians and datasets.

5.5 Comparison to Automated Techniques

As a point of comparison, we also used purely automated techniques to classify the same dataset from the preliminary study. This comparison was made possible because Historian A had labeled all 189 documents used in the preliminary study with his research topics (see Table 5). Note that the comparison conditions are not fully equivalent, as the crowd required far less data. The crowd had only two relevant examples, whereas the following machine learning models need both relevant and irrelevant examples. Also, to maximize the power of automated techniques, we used all 189 documents with the four topics, compared to only the two example documents needed for RAP.

Since all the primary sources are digitized and in an image format, our first step was to use an optical character recognition (OCR) system, Tesseract 4.00.00a (with LSTM)[2] (R. Smith, 2007; R. W. Smith, 2009; Ray Smith, Antonova, & Lee, 2009), to automatically transcribe them. Next, we preprocessed these textual documents by removing stopwords and stemming words based on a Snowball algorithm. We then transformed these preprocessed documents into TF-IDF space. Next, we chose five techniques representing five different categories of algorithms for binary text classification: 1) logistic regression, 2) kNN (k=9 to maximize available class samples), 3) SVM, 4) decision tree (CART), and 5) random forest (Sebastiani, 2002). Finally, we ran stratified 10-fold cross validations for all five techniques for each of the historian's four topics.

The results show that all techniques have high accuracy (0.75-0.95) but very low recall (0-0.3) due to the highly imbalanced numbers of class samples. For example, there were only 10 out of 189 documents relevant to *American Hypocrisy* for which accuracy is 0.91-0.95, but recall is 0 across all techniques. This means all relevant documents were missed for that topic.

To deal with the class imbalance issue, we applied three common techniques: adjusting class weights, random over-sampling, and random under-sampling. Under-sampling showed the best improvement, with accuracy 0.55-0.8 and recall 0.2-0.7. For example, with under-sampling, SVM had highest recall (0.7) and 0.55 accuracy for *American Hypocrisy*. Although this was a substantial improvement in recall, it may still not be practical, because there were very few relevant examples for this topic, and 30% of the relevant documents were mistakenly excluded by the automated technique.

Although future advancements may make automated techniques more powerful, the above results show RAP and crowdsourcing may offer a compelling alternative in our demonstrated context of history.

2 https://github.com/tesseract-ocr/tesseract

6. BROADER IMPLICATIONS

6.1 Quality Labels for Historical Scholarship

The Read-Agree-Predict (RAP) algorithm allows novice transient crowds to find relevant primary sources in a digital archive as effectively as expert historians and, as a byproduct, reveals and prioritizes crowds' confusions. We demonstrated the effectiveness of this approach with an authentic historical dataset and two studies with different historians, topics of interest, and documents. RAP also offers clear advantages over majority vote. Our empirical results and simulations show that RAP consistently outperforms majority vote, and larger crowd sizes increase RAP's accuracy to be on par with experts. We also propose that a major strength of RAP is that its design is simple and elegant enough to be easily implemented for a variety of systems.

The ability to produce quality labels may give more confidence to historians in trusting data collected via crowdsourcing and in adopting this new crowdsourcing model for their research and classes. By doing so, historians can focus more on exploring interesting research questions and using relevant sources to support their arguments. Anecdotally, our experts Historian A and Historian B were excited to see how crowds could help with their research. Asked about his interest in using RAP for crowdsourced support of archival research, Historian A was enthusiastic:

> *"Definitely! Yeah, I mean that'll be very useful. It's often kind of difficult to do, especially with a topic like nationalism, because it would be hard to just do a keyword search, because nobody was using the word 'nationalism' in the 1860s. So, to some extent, you just have to read, you know, everything in them [digital archives] and kind of hope something useful comes along."*

6.2 Opportunities for History Education

Historical primary sources are important sources for both scholarly research and education in history domain (Stearns et al., 2000; Tally & Goldenberg, 2005) and teaching students to "think like a historian" is a primary goal of history education (Hynd, Holschuh, & Hubbard, 2004; Mandell, 2008; Wineburg, 2010). While the experiments in this article focus on paid crowd workers as a proof-of-concept, future work may explore how RAP extends to students in classroom settings. Within this context, RAP's crowdsourcing model may create a win-win situation for both historian-educators and students. One the one hand, this model could help historians do research by organizing related primary sources into research topics, and to teach by identifying and prioritizing students' confusions. On the other hand, students could get opportunities to participate in authentic historical research and to practice historical thinking and knowledge with primary sources and receive feedback accordingly. As prior research shows, comparing students' and domain experts' output of the same task is an effective way to identify

students' confusions (Anderson, Boyle, Corbett, & Lewis, 1990; Anderson, Boyle, & Reiser, 1985; Merrill, Reiser, Ranney, & Trafton, 1992).

By adopting the new crowdsourcing model and RAP in classroom settings, historians could easily organize unprocessed primary sources and collect prioritized confusions that may be pervasive among students, and direct their time and expertise to the ones with higher potential impact. Demystified materials may help motivate and engage students, as research shows that people are often interested in surprising materials that challenge their existing assumptions (Brands, 2008; Davis, 1971).

While other research from non-historical domains shows that it is possible for the crowd to learn a few microtasks in a short amount of time (e.g., < 30 minutes in total) (Dow, Kulkarni, Klemmer, & Hartmann, 2012; Lee, Lo, Kim, & Paulos, 2016; Zhu, Dow, Kraut, & Kittur, 2014), we did not observe this in our study of reading comprehension techniques. The wide adoption of long-term apprenticeship in historical research may help explain why we have different results (Law et al., 2017).

6.2.1 Classroom Usage Scenario

To make concrete the potential costs and benefits of deploying RAP in a classroom setting, we propose the following usage scenario. A historian could begin with a list of topics of interests and a collection of unprocessed primary sources. In the historian's class, she picks topics of interest relevant to the class and asks students to apply relevance labels for these topics to unprocessed primary sources. As the class progresses, the RAP automatically reports relevant primary sources for the topics and prioritizes students' confusions about sources and topics by aggregating students' labels. The historian then uses relevant sources for research and addresses students' confusions starting with the prioritized errors.

Taking data from the preliminary study as an example, with 189 primary sources and 4 topics relevant to the class, it would take 9828 human minutes to complete all possible labels (189 sources × 4 topics × 5 students per source-topic pair × 2.6 average reading time per source). Historian A generally has about 35 students in his course on the American Civil War, and there are about 16 weeks per semester, requiring only about 17 minutes per week for each student in a semester. In practice, students should be able to analyze more sources as they learn to improve their skills throughout the process, and five students are not always needed when there is already high agreement.

Assuming the historian's processing time is roughly the same, it would require the historian to spend 1965.6 minutes (32.8 hours) to analyze 189 sources (189 sources × 4 topics × 1 historian per source-topic pair × 2.6 average processing time per source). Thus, the historian could save nearly 33 hours of analysis time by leveraging RAP within a classroom context, excluding time

spent on developing initial examples and monitoring students, which to some extent may overlap with existing teaching responsibilities.

Alternatively, if the historian were to use RAP with paid crowd workers, the total cost for labeling 189 documents for 4 topics would be about $1187.60 (9828 minutes / 60 minutes × $7.25/hour) using the current US Federal minimum wage, or about $1.57 per document-topic analysis. While using paid crowds requires financial resources, this approach has the advantage of being much faster (completable within a day) than the semester-long classroom-based scenario described above. RAP's high-quality results in our crowd-based studies suggest a variety of options to historians based on their available time, funding, and teaching flexibility.

6.3 Limitations and Future Work

This article focused on understanding baseline use of crowdsourcing in historical scholarship and reported several studies with real-world digital archives. Our findings are based on a set of five historical topics and 18 primary sources, averaging 250 words in length and at a college-age reading level, from the American Civil War era. Additional studies, drawing on larger datasets of topics and documents, exploring other historical periods and document formats, and adapting the techniques for classroom settings, are needed to show how these findings replicate and generalize. Experts in many other domains also serve as both researchers and educators, so we believe this new crowdsourcing model may also apply to other domains. Furthermore, the notion of labeling raw textual documents with high-level concepts is prevalent in sensemaking tasks such as intelligence analysis (Pirolli & Card, 2005) and software or product design (Russell, Stefik, Pirolli, & Card, 1993), so we believe the RAP will generalize to other related domains.

To understand the baseline performance with minimal constraints, our studies used novice paid crowds. With these positive initial results, we have more confidence to deploy the new crowdsourcing model in a real classroom setting as the next step. In a classroom setting, we might want to include work history, so we can apply more automated techniques such as EM to further increase robustness of RAP, decrease the required participants per document, or both. In addition, work history can also double as learning history to constantly reflect how well students learn throughout the process.

7. CONCLUSION

With digitized historical and scholarly materials made available online, it is often difficult for researchers to find documents of interest because the topics and themes they are investigating are specialized and abstract. In this article, we investigated the possibility of a new crowdsourcing model to label the relevance of digitized primary sources to high-level topics, and to reveal and prioritize crowd confusions. In our preliminary study, focusing on the effect of different semantic tasks on comprehension, we found promising results supporting the new crowdsourcing model. We also found that a robust pattern emerged enabling highly accurate predictions of document

relevance based on crowd performance. Based on these results, we developed Read-Agree-Predict (RAP), a crowdsourcing approach which allows crowds to label relevance of primary sources to an abstract theme with high accuracy. As a useful byproduct, RAP also reveals crowd confusions that suggest opportunities for learning interventions. We successfully validated RAP with a new historian and set of primary sources, and conducted follow-up analyses with a simulation study and a comparison of agreement among experts. While this research used paid crowd workers in a historical domain, it has implications for applications in classroom settings and in other domains.

8. ACKNOWLEDGEMENTS

We wish to thank Daniel Newcomb and our MTurk study participants. This research was supported by U.S. National Historical Publications and Records Commission Grant DH50013-15.

9. REFERENCES

Aggarwal, C. C., & Zhai, C. (2012). A Survey of Text Classification Algorithms. In C. C. Aggarwal & C. Zhai (Eds.), Mining Text Data (pp. 163–222). https://doi.org/10.1007/978-1-4614-3223-4_6

Anderson, J. R., Boyle, C. F., Corbett, A. T., & Lewis, M. W. (1990). Cognitive modeling and intelligent tutoring. Artificial Intelligence, 42(1), 7–49.

Anderson, J. R., Boyle, C. F., & Reiser, B. J. (1985). Intelligent tutoring systems. Science(Washington), 228(4698), 456–462.

André, P., Kittur, A., & Dow, S. P. (2014). Crowd Synthesis: Extracting Categories and Clusters from Complex Data. Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, 989–998. https://doi.org/10.1145/2531602.2531653

Banko, M., & Brill, E. (2001). Scaling to Very Very Large Corpora for Natural Language Disambiguation. Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, 26–33. https://doi.org/10.3115/1073012.1073017

Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., … Panovich, K. (2010). Soylent: A Word Processor with a Crowd Inside. Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology, 313–322. https://doi.org/10.1145/1866029.1866078

Bobrow, S. A., & Bower, G. H. (1969). Comprehension and recall of sentences. Journal of Experimental Psychology, 80(3, Pt.1), 455–461. https://doi.org/10.1037/h0027461

Brands, H. W. (2008). Response to Hochschild. Historically Speaking, 9(4), 6–7. https://doi.org/10.1353/hsp.2008.0063

Bretzing, B. H., & Kulhavy, R. W. (1979). Notetaking and depth of processing. Contemporary Educational Psychology, 4(2), 145–153. https://doi.org/10.1016/0361-476X(79)90069-9

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. The Journal of the Learning Sciences, 2(2), 141–178.

Cai, C. J., Iqbal, S. T., & Teevan, J. (2016). Chain Reactions: The Impact of Order on Microtask Chains. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 3143–3154. https://doi.org/10.1145/2858036.2858237

Chi, E. H., Hong, L., Heiser, J., & Card, S. K. (2006). Scentindex: Conceptually Reorganizing Subject Indexes for

Reading. 2006 IEEE Symposium On Visual Analytics Science And Technology, 159–166. https://doi.org/10.1109/VAST.2006.261418

Chi, Ed H., Hong, L., Gumbrecht, M., & Card, S. K. (2005). ScentHighlights: Highlighting Conceptually-related Sentences During Reading. Proceedings of the 10th International Conference on Intelligent User Interfaces, 272–274. https://doi.org/10.1145/1040830.1040895

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. Journal of Verbal Learning and Verbal Behavior, 11(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. Journal of Experimental Psychology: General, 104(3), 268–294. https://doi.org/10.1037/0096-3445.104.3.268

Dalton, M. S., & Charnigo, L. (2004). Historians and their information sources. College & Research Libraries, 65(5), 400–425.

Davis, M. S. (1971). That's Interesting: Towards a Phenomenology of Sociology and a Sociology of Phenomenology. Philosophy of the Social Sciences, 1(4), 309–344.

Dawid, A. P., & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. Applied Statistics, 20–28.

Doctorow, M., C, M., & Marks, C. (1978). Generative processes in reading comprehension. Journal of Educational Psychology, 70(2), 109–118. https://doi.org/10.1037/0022-0663.70.2.109

Dow, S., Kulkarni, A., Klemmer, S., & Hartmann, B. (2012). Shepherding the Crowd Yields Better Work. Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, 1013–1022. https://doi.org/10.1145/2145204.2145355

Drapeau, R., Chilton, L. B., Bragg, J., & Weld, D. S. (2016). MicroTalk: Using Argumentation to Improve Crowdsourcing Accuracy. Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP). Retrieved from http://www.cs.washington.edu/ai/pubs/drapeau-hcomp16.pdf

Duan, X., & Tajima, K. (2019). Improving Multiclass Classification in Crowdsourcing by Using Hierarchical Schemes. The World Wide Web Conference, 2694–2700. ACM.

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). Statistical methods for rates and proportions. John Wiley & Sons.

Giancola, M., Paffenroth, R., & Whitehill, J. (2018). Permutation-invariant consensus over crowdsourced labels. Sixth AAAI Conference on Human Computation and Crowdsourcing.

Glassman, E. L., Kim, J., Monroy-Hernández, A., & Morris, M. R. (2015). Mudslide: A Spatially Anchored Census of Student Confusion for Online Lecture Videos. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 1555–1564. https://doi.org/10.1145/2702123.2702304

Glassman, E. L., Lin, A., Cai, C. J., & Miller, R. C. (2016). Learnersourcing Personalized Hints. Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, 1626–1636. https://doi.org/10.1145/2818048.2820011

Grant, S. G. (2001). It's just the facts, or is it? The relationship between teachers' practices and students' understandings of history. Theory & Research in Social Education, 29(1), 65–108.

Grant, S. G. (2018). Teaching Practices in History Education. In The Wiley International Handbook of History Teaching and Learning (pp. 419–448). https://doi.org/10.1002/9781119100812.ch16

Grant, S. G., & Gradwell, J. M. (2010). Teaching History with Big Ideas: Cases of Ambitious Teachers. R&L Education.

Hosseini, M., Cox, I. J., Milić-Frayling, N., Kazai, G., & Vinay, V. (2012). On Aggregating Labels from Multiple Crowd Workers to Infer Relevance of Documents. Advances in Information Retrieval, 182–194. https://doi.org/10.1007/978-3-642-28997-2_16

Hynd, C., Holschuh, J. P., & Hubbard, B. P. (2004). Thinking like a historian: College students' reading of multiple

historical documents. Journal of Literacy Research, 36(2), 141–176.

Ipeirotis, P. G., Provost, F., & Wang, J. (2010). Quality Management on Amazon Mechanical Turk. Proceedings of the ACM SIGKDD Workshop on Human Computation, 64–67. https://doi.org/10.1145/1837885.1837906

John Uebersax. (2009). Raw Agreement Indices. Retrieved August 8, 2019, from http://www.john-uebersax.com/stat/raw.htm

Kavzoglu, T., & Colkesen, I. (2012). The effects of training set size for performance of support vector machines and decision trees. Proceeding of the 10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, July, 10–13.

Kim, J., Miller, R. C., & Gajos, K. Z. (2013). Learnersourcing Subgoal Labeling to Support Learning from How-to Videos. CHI '13 Extended Abstracts on Human Factors in Computing Systems, 685–690. https://doi.org/10.1145/2468356.2468477

Kim, J., & others. (2015). Learnersourcing: Improving learning with collective learner activity (Massachusetts Institute of Technology). Retrieved from http://dspace.mit.edu/handle/1721.1/101464

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. Psychological Review, 85(5), 363–394. https://doi.org/10.1037/0033-295X.85.5.363

Law, E., Gajos, K. Z., Wiggins, A., Gray, M. L., & Williams, A. (2017). Crowdsourcing As a Tool for Research: Implications of Uncertainty. Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 1544–1561. https://doi.org/10.1145/2998181.2998197

Lee, D. J., Lo, J., Kim, M., & Paulos, E. (2016). Crowdclass: Designing classification-based citizen science learning modules. Retrieved from http://dorisjunglinlee.com/files/crowdclass.pdf

Linden, M., & Wittrock, M. C. (1981). The Teaching of Reading Comprehension according to the Model of Generative Learning. Reading Research Quarterly, 17(1), 44–57. https://doi.org/10.2307/747248

Little, G., Chilton, L. B., Goldman, M., & Miller, R. C. (2010). Exploring Iterative and Parallel Human Computation Processes. Proceedings of the ACM SIGKDD Workshop on Human Computation, 68–76. https://doi.org/10.1145/1837885.1837907

Mandell, N. (2008). Thinking like a Historian: A Framework for Teaching and Learning. OAH Magazine of History, 22(2), 55–59. https://doi.org/10.1093/maghis/22.2.55

McDaniel, K. N. (2010). Harry Potter and the Ghost Teacher: Resurrecting the Lost Art of Lecturing. The History Teacher, 43(2), 289–295. Retrieved from JSTOR.

McDonnell, T., Lease, M., Elsayad, T., & Kutlu, M. (2016). Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments. Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP). Retrieved from https://www.ischool.utexas.edu/~ml/papers/mcdonnell-hcomp16.pdf

Merrill, D. C., Reiser, B. J., Ranney, M., & Trafton, J. G. (1992). Effective Tutoring Techniques: A Comparison of Human Tutors and Intelligent Tutoring Systems. Journal of the Learning Sciences, 2(3), 277–305. https://doi.org/10.1207/s15327809jls0203_2

Mitros, P. (2015). Learnersourcing of Complex Assessments. Proceedings of the Second (2015) ACM Conference on Learning @ Scale, 317–320. https://doi.org/10.1145/2724660.2728683

Nawrotzki, K. (Ed.). (2013). Writing History in the Digital Age. Retrieved from http://hdl.handle.net/2027/spo.12230987.0001.001

Nist, S. L., & Hogrebe, M. C. (1987). The Role of Underlining and Annotating in Remembering Textual Information. Reading Research and Instruction, 27(1), 12–25. https://doi.org/10.1080/19388078709557922

Peterson, S. E. (1991). The cognitive functions of underlining as a study technique. Reading Research and Instruction, 31(2), 49–56. https://doi.org/10.1080/19388079209558078

Pirolli, P., & Card, S. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. Proceedings of International Conference on Intelligence Analysis, 5, 2–4. Retrieved from https://www.e-education.psu.edu/geog885/sites/www.e-education.psu.edu.geog885/files/geog885q/file/Lesson_02/Sense_Making_206_Camera_Ready_Paper.pdf

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. Nature, 541(7638), 532–535. https://doi.org/10.1038/nature21054

Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993). The Cost Structure of Sensemaking. Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems, 269–276. https://doi.org/10.1145/169059.169209

Rutner, J., & Schonfeld, R. (2012). Supporting the Changing Research Practices of Historians. Retrieved from Ithaka S+R website: http://sr.ithaka.org/?p=22532

Schnell, T., & Rocchio, D. (1978). A Comparison of Underlying Strategies for Improving Reading Comprehension and Retention. Reading Horizons, 18(2). Retrieved from http://scholarworks.wmich.edu/reading_horizons/vol18/iss2/4

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. ACM Comput. Surv., 34(1), 1–47. https://doi.org/10.1145/505282.505283

Sheng, V. S., Provost, F., & Ipeirotis, P. G. (2008). Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 614–622. https://doi.org/10.1145/1401890.1401965

Šimko, J., Šimko, M., Bieliková, M., Ševcech, J., & Burger, R. (2013). Classsourcing: Crowd-Based Validation of Question-Answer Learning Objects. International Conference on Computational Collective Intelligence, 62–71. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-40495-5_7

Smart, K. L., & Bruning, J. L. (1973). An examination of the practical importance of the von Restorff effect. Annual Meeting of the American Psychological Association, Montreal, Canada.

Smith, R. (2007). An Overview of the Tesseract OCR Engine. Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 2, 629–633. https://doi.org/10.1109/ICDAR.2007.4376991

Smith, R. W. (2009). Hybrid Page Layout Analysis via Tab-Stop Detection. Proceedings of the 2009 10th International Conference on Document Analysis and Recognition, 241–245. https://doi.org/10.1109/ICDAR.2009.257

Smith, Ray, Antonova, D., & Lee, D.-S. (2009). Adapting the Tesseract Open Source OCR Engine for Multilingual OCR. Proceedings of the International Workshop on Multilingual OCR, 1:1–1:8. https://doi.org/10.1145/1577802.1577804

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and Fast—But is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 254–263. Retrieved from http://dl.acm.org/citation.cfm?id=1613715.1613751

Stearns, P. N., Seixas, P. C., & Wineburg, S. (2000). Knowing, teaching, and learning history: National and international perspectives. Retrieved from https://books.google.com/books?hl=en&lr=&id=viQVCgAAQBAJ&oi=fnd&pg=PR9&dq=+Knowing,teaching,+and+learning+history&ots=gPjNC0qroE&sig=RxJx6hzT9Cq0-CTOdyk8RhDbTBs

Tally, B., & Goldenberg, L. B. (2005). Fostering historical thinking with digitized primary sources. Journal of Research on Technology in Education, 38(1), 1–21.

Venkatesan, R., Er, M. J., Dave, M., Pratama, M., & Wu, S. (2016). A novel online multi-label classifier for high-speed streaming data applications. Evolving Systems, 1–13.

Weir, S., Kim, J., Gajos, K. Z., & Miller, R. C. (2015). Learnersourcing Subgoal Labels for How-to Videos. Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, 405–416. https://doi.org/10.1145/2675133.2675219

Wineburg, S. (2010). Thinking like a historian. Teaching with Primary Sources Quarterly, 3(1), 2–4.

Wittrock, M. C., & Alesandrini, K. (1990). Generation of Summaries and Analogies and Analytic and Holistic Abilities. American Educational Research Journal, 27(3), 489–502. https://doi.org/10.3102/00028312027003489

Wittrock, Merlin C. (1989). Generative Processes of Comprehension. Educational Psychologist, 24(4), 345–376. https://doi.org/10.1207/s15326985ep2404_2

Xu, A., Rao, H., Dow, S. P., & Bailey, B. P. (2015). A Classroom Study of Using Crowd Feedback in the Iterative Design Process. Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, 1637–1648. https://doi.org/10.1145/2675133.2675140

Yu, L., Kittur, A., & Kraut, R. E. (2014). Distributed Analogical Idea Generation: Inventing with Crowds. Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems, 1245–1254. https://doi.org/10.1145/2556288.2557371

Zhang, M.-L., & Zhou, Z.-H. (2014). A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering, 26(8), 1819–1837.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. Advances in Neural Information Processing Systems, 649–657. Retrieved from http://papers.nips.cc/paper/5782-character-level-convolutional-networks-for-text-classification

Zhu, H., Dow, S. P., Kraut, R. E., & Kittur, A. (2014). Reviewing Versus Doing: Learning and Performance in Crowd Assessment. Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, 1445–1455. https://doi.org/10.1145/2531602.2531718