Evaluating Topic-Word Review Analysis for Understanding Student Peer Review Performance

Wenting Xiong  
University of Pittsburgh  
Pittsburgh, PA, 15260  
wex12@cs.pitt.edu

Diane Litman  
University of Pittsburgh  
Pittsburgh, PA, 15260  
litman@cs.pitt.edu

ABSTRACT
Topic modeling is widely used for content analysis of textual documents. While the mined topic terms are considered as a semantic abstraction of the original text, few people evaluate the accuracy of humans’ interpretation of them in the context of an application based on the topic terms. Previously, we proposed RevExplore, an interactive peer-review analytic tool that supports teachers in making sense of large volumes of student peer reviews. To better evaluate the functionality of RevExplore, in this paper we take a closer look at its Natural Language Processing component which automatically compares two groups of reviews at the topic-word level. We employ a user study to evaluate our topic extraction method, as well as the topic-word analysis approach in the context of educational peer-review analysis. Our results show that the proposed method is better than a baseline in terms of capturing student reviewing/writing performance. While users generally identify student writing/reviewing performance correctly, participants who have prior teaching or peer-review experience tend to have better performance on our review exploration tasks, as well as higher satisfaction towards the proposed review analysis approach.

Keywords  
Educational peer reviews, text analysis, topic modeling, user study

1. INTRODUCTION
Peer review is a popular educational approach for helping students improve their writing performance. It provides different perspectives and valuable feedback on what is compelling and what is problematic. Ideally, from analyzing student peer reviews, instructors may not only learn about student writing issues by reading student feedback, but may also evaluate student reviewing performance by checking if comments are given for important issues in a good manner. However, due to the large amount of reviews, teachers seldom read the comments carefully if at all. Instructors whom we have interviewed have complained that peer reviews are time consuming to read and difficult to interpret. Interpreting them requires synthesizing opinions from multiple parties while making comparisons and contrasts across multiple students at the same time.

Nowadays, some existing web-based peer-review systems can help teachers set up peer review assignments and even grade student papers based on peer ratings, though no software yet has the intelligence to support teachers’ comprehension of the textual review comments. Previously [14], we took our first step to address this issue and designed an interactive analytic interface (RevExplore) on top of SWoRD [4], a web-based peer-review reciprocal system that has been used by over 12,000 students over the last 8 years. Before deploying RevExplore as a SWoRD plugin to the public, we would like to evaluate its functionality carefully, especially its natural language processing (NLP) component that automatically abstracts and compares review content at the topic-word level.

For this purpose, we carry out a user study to examine the idea of analyzing peer reviews by comparing them in groups based on their topic words. In particular, we investigate the analytic power of topic words in the context of assessing student writing/review performance by mining peer reviews. In this study, we not only show that our proposed topic-word extraction method can better enable users to identify student writing/reviewing issues than a baseline, but also demonstrate that the utility of our topic-word approach depends on various factors.

2. RELATED WORK
There is increasing interest in research on computer-supported peer reviews both from the students’ perspective for improving learning and from the teachers’ perspective for informing decision making. From the students’ perspective, prior studies of automatically assessing student peer-review performance either focus on detecting important feedback features [3, 15], or aim to assess the overall peer-review helpfulness [13]. From the teachers’ perspective, Goldin and Ashley [6] use Bayesian networks to model computer-supported peer review which yields pedagogically useful information about student learning and about grading schema. In contrast with their work, we are interested in the educational contents (textual peer reviews) rather than the interaction between students during the peer review activities. Further-
more, our RevExplore involves humans in the loop: it allows teachers to interactively explore peer-review data at the student level first, and then drill down to particular groups of students for automated analysis of their peer reviews afterwards.

With respect to data mining of educational textual contents, the general goal is to summarize or analyze the textual contents to provide feedback to teachers, either about student learning activities, or about the utility of teaching materials. For understanding student learning activities, word-based content analysis, where the words are either learned through topic modeling or crafted manually, has been widely used for categorizing educational data such as online discussion threads [10, 9] and student-tutor interactions [7]. As peer review provides students learning opportunities during both paper writing and reviewing, the textual reviews are valuable in reflecting both student writing and reviewing performance. Therefore we hypothesize that extending data mining techniques to textual peer reviews can provide useful feedback to instructors regarding both student writing and reviewing performance.

Several NLP techniques can be used for word-based content analysis. One is the frequency-based method, which considers the content of a target corpus in terms of its most frequent words. A famous application of this method is to generate word-clouds, which is a popular web2.0 tool for supporting impression formation over textual data. For example, it has been used to compare political speeches from different people. In our study, we consider this method as a baseline (denoted as \( F_{\text{req}} \)) for evaluating our proposed topic-word extraction method, which is to automatically learn the salient words of a target corpus through a topic-signature approach to topic modeling (denoted as TopicS). Topic signature modeling assumes a single topic of the target corpus when comparing it against a background corpus. And this topic can be represented as a set of words based on statistical analysis of the word distribution in both corpora [8].

Another kind of topic modeling is based on graphical models, such as LDA [1]. LDA considers each document as a mixture over an underlying set of topic probabilities. While it has been widely studied for many NLP tasks from sentiment analysis to text summarization, we did not employ it in RevExplore for several reasons. First, the learned topic model changes with parameter settings (the number of topics and the hyperparameters) which are quite task dependent. Furthermore, the learned topics are generally difficult to interpret [16], and hard to evaluate. Although various automatic metrics were proposed, they do not always agree with human judgements in end-applications [2].

3. TOPIC WORDS IN REVEXPLORE

RevExplore [14] utilizes data visualization in combination with NLP techniques to help instructors interactively make sense of peer review data, which was almost impractical before. It has a student performance overview and a review comparison detail-view. In the overview, RevExplore visualizes the overall peer-review information at the student level, which allows instructors to effectively identify points of interest during their initial data exploration. In the detail-view, RevExplore automatically abstracts the semantic information of peer reviews at the topic-word level with the original texts visible on demand. To create the detail-view, we adapt existing natural language processing techniques to the peer-review domain for supporting automated analytics.

3.1 Preprocessing – domain word masking

Because peer reviews frequently refer to the content of the papers that they comment on, it is necessary to reduce the influence of such “paper topic” words on the extraction of “review topic” words from the peer reviews, otherwise the “paper topic” will dominate the computation of “review topic” words. Therefore, as a preprocessing step, we first compute the “paper topic” words of the writing assignment using TopicS, a java implementation of the topic signature acquisition algorithm [8]. TopicS computes the topic words from a topic relevant (target) corpus against a topic irrelevant (background) corpus based on word distribution using chi-square statistics (which will be explained later in this section). For computing the “paper topic” words, we use all student papers as the target corpus and use 5000 documents from the English Gigaword Corpus as the background corpus (the default setting of TopicS). Based on our intuition, we set the chi-square cutoff to be 10 (\( p = .0016 \)), yielding about 500 topic words. As these words depend on the domain of the writing assignment, we denote them as domain words for the rest of the paper.

To prevent analysts from being distracted too much by domain words when analyzing peer reviews, before computing the “review topic” words using any extraction method, we apply domain word masking to each peer review by replacing all occurrences of each domain word (e.g. “war”, “african”, “americans”, “women”, “democracy”, “rights”, “states”) with a dummy term “domainwords”.

3.2 Comparison-oriented topic signatures

The topic signature algorithm [8] assumes that a target corpus has a single topic, and it computes the topic words for the target corpus with respect to a general background corpus. For each word, the algorithm computes a likelihood ratio [5] which tests the hypothesis of the word being a topic word of the target corpus versus the hypothesis that the word is not a topic word. The −2 log likelihood ratio has a chi-square distribution, which allows us to test the significance of each word to the topic of the target corpus when compared against the background corpus. In our work, we use the existing software TopicS (as mentioned before) for extracting topic signatures.

In RevExplore overview, when two groups of students are chosen for further review comparison in the detail-view, there is an implicit assumption of a topic difference between their corresponding review groups. Furthermore, the topic to be mined changes dynamically in accordance with the change of the analytic goals, which are specified through different grouping of reviews. To capture these assumptions when using TopicS to extract the topic words for a particular review group, we take its reviews as the target corpus and use all of

\[^1\]http://www.tagcrowd.com/blog/2011/03/05/state-of-the-union-2002-vs-2011/

\[^2\]TopicS was developed by Anni Louis for evaluating automated text summarization [12].
the reviews as the background corpus. In this way, we tailor the computation of the topic words to the desired analytic property of the target review group. We denote this adapted method as TopicS. For TopicS, we set the significance cutoff as 6.635, corresponding to a p value of .01. For a typical review group of our study, the number of the extracted topic words is about 20.

In our user study, we compare TopicS with the frequency-based method (Freq), and we expect that our TopicS can outperform Freq in helping users achieve better task performance. Note that both methods are performed after the domain-word masking.

4. DATA
Our peer-review corpus consists of 1405 free-text review comments and 24 student papers, which were collected in a college level history class [11]. The peer review was done through SWoRD [4], a web-based peer-review reciprocal system, as follows:

Assignment creation: The teacher first created the writing assignment in SWoRD and provided a peer-review rubric which required students to assess a paper’s quality on three separate dimensions (Logic, Flow and Insight), by giving a numeric rating on a scale of 1-7 in addition to textual comments.3 For instance, the teacher created the following guidance for commenting on the “Logic” dimension: “Provide specific comments about the logic of the author’s argument. If points were just made without support, describe which ones they were. If the support provided doesn’t make logical sense, explain what that is. If some obvious counter-argument was not considered, ...”. Teacher guidance for numerically rating the logical arguments of the paper was also given. For this history assignment, a rating of 7 (“Excellent”) was described as “All arguments strongly supported and no logical flaws in the arguments.”. A rating of 1 (“Disastrous”) was described as “No support presented for any arguments, or obvious flaws in all arguments.”. Textual review examples for the Flow dimension are provided in Figure 1 of Section 6.

Paper writing & peer review: In the next phase, 24 students submitted their papers online through SWoRD and then reviewed 6 peers’ papers. The peer review was done in a “double blind” manner and each paper was reviewed by about 6 peers. As students were required to submit reviews on each dimension separately, SWoRD automatically associates the reviewing dimension with every numerical rating and textual comment. In addition, students also received reviews from one content expert and another writing expert, who reviewed in the same way as the peers did, yielding a final 1405 review comments4 that we use in this study. In our user study, we will group students based on their writing performance as determined by the numerical peer review ratings. In particular, the average of peers’ paper ratings received by each student (ratingW) measures the overall quality of a student’s writing performance.

Backward evaluation: Finally, peer feedback was rated

3While reviewing dimensions and associated rubrics are typically created by the teacher, teachers can also use a library provided by SWoRD.

4A single peer review can have multiple comments.
three questions in the user study, e.g. for analyzing student writing performance:

- Q1: Considering students who received the reviews, which group do you think might be labeled as “high” in terms of their writing performance?
- Q2: Within one minute or two, can you figure out how one group of reviews focus on different issues/aspects/scope compared to the other?
- Q3: Comparing the two topic extraction methods, given the correct labels, which method is more helpful in discovering the group difference of reviewing focus and content?

For user satisfaction analysis, we would like to know how the utility of the proposed idea is affected by user background information, especially participant prior experience of peer review and teaching. We examine these factors in terms of both users’ task performance and their reported satisfaction (subjective ratings) in an exit survey.

**Participants:** All 46 participants are students recruited from a university campus, who are from various academic backgrounds including English, Linguistics, Psychology, Education, Computer Science, etc. Although the tool is designed for instructors, it is quite difficult to recruit a sizeable number and thus we recruit students instead. Note that some students do have teaching experience and are experienced SWoRD users. To understand whether such background plays a significant role in the use of RevExplore, we record user background information especially regarding demographic factors that depict participants’ prior experience in peer review and teaching: whether they have peer-review experience before (expPR), whether they used SWoRD before (expSWoRD), whether they were a TA before (expTA), and whether they have graded any writing assignment before (expGW). Participant distribution over these factors is presented in Table 1. Although we also look at other demographic factors such as age, gender, major, etc., due to the space limit, we do not report them in this paper.

**Procedure:** Before being exposed to the analysis tasks, participants were first given instructions about the peer-review assignment, including both the paper topics and the reviewing rubrics. We also provided a warm up example to demonstrate how to analyze peer reviews through our user study interface. Figure 1 is a screenshot of the interface, which consists of three parts: the left pane displays the original reviews in lowercase after removing non-ascii characters; the middle pane shows the topic words extracted from the two groups of reviews; the right pane shows the analysis.

![Figure 1: Interface annotation for peer review analysis user study.](image-url)
Table 2: Descriptive statistics of user satisfaction. Higher rating means more positive opinion except for Q_textRef and Q_textImp. One sample t-test test value = 3 (neutral). Significant items are highlighted in bold ($p < 0.05$).

<table>
<thead>
<tr>
<th>Question</th>
<th>Content</th>
<th>Mean</th>
<th>Std.Error</th>
<th>Sig.(2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q_easyNess</td>
<td>Is it easy to make sense of reviews by comparing topic words?</td>
<td>2.85</td>
<td>.140</td>
<td>.291</td>
</tr>
<tr>
<td>Q_listDiff</td>
<td>Do the two lists of topic words from the two review groups look semantically different to you?</td>
<td>3.52</td>
<td>.123</td>
<td>.000</td>
</tr>
<tr>
<td>Q_layout</td>
<td>What do you think of the list-layout of topic words for comparison purpose?</td>
<td>3.57</td>
<td>.154</td>
<td>.001</td>
</tr>
<tr>
<td>Q_reviewDiff</td>
<td>What do you think of topic comparison in helping you identify the differences in the peer reviews?</td>
<td>3.54</td>
<td>.145</td>
<td>.000</td>
</tr>
<tr>
<td>Q_largeData</td>
<td>What do you think of topic comparison in helping you make sense of large amount of peer reviews?</td>
<td>3.93</td>
<td>.177</td>
<td>.000</td>
</tr>
<tr>
<td>Q_approach</td>
<td>How do you like the idea of exploring peer reviews by comparing them in groups using their topic words? (comparing to reading the textual reviews?)</td>
<td>3.46</td>
<td>.180</td>
<td>.015</td>
</tr>
<tr>
<td>Q_textRef</td>
<td>How often did you refer to the original reviews to make sense of the topic words?</td>
<td>1.96</td>
<td>.189</td>
<td>.000</td>
</tr>
<tr>
<td>Q_textImp</td>
<td>How important is the original reviews for you to analyze the group differences?</td>
<td>2.93</td>
<td>.171</td>
<td>.705</td>
</tr>
</tbody>
</table>

7. EXPERIMENT RESULTS

The statistics of the task performance are summarized in Table 3. For measuring task performance, we use the following scheme to code participants’ answers to the three questions:

\[
\text{Answer}_1 = \begin{cases} 
1 & \text{if the answer is correct,} \\
-1 & \text{if the answer is incorrect,} \\
0 & \text{if "I have no idea".}
\end{cases} 
\]

(1)

\[
\text{Answer}_2 = \begin{cases} 
1 & \text{if yes,} \\
0 & \text{if "no".}
\end{cases} 
\]

(2)

\[
\text{Answer}_3 = \begin{cases} 
1 & \text{if vote for TopicS,} \\
-1 & \text{if vote for Freq,} \\
0 & \text{if "no preference".}
\end{cases} 
\]

(3)

For each question, we compare participants’ answers to random guess using a one sample t-test to check if using the topic words (extracted by either method) is generally meaningful for our peer-review analysis tasks. As the table shows, in general, the proposed approach is better than random guess (the corresponding test mean is: 0, 0.5, 0), and the proposed topic extraction method (TopicS) yields better task performance than the baseline (Freq). However, we also notice that the task performance varies with the analysis tasks. This motivates us to further examine the effects of Split and Dim, as well as their interaction with Method, which is discussed later.

To analyze user satisfaction, we compare participants’ rating of each survey item to the neutral state (3-point) using a one sample t-test. As Table 2 shows, despite that participants generally think the analysis task is neither easy or difficult,
Table 3: Summary of estimates of the variables across all different conditions, with higher mean bolded between the two extraction methods. It shows that TopicS generally yields higher mean compared with Freq, except for predicting the label of topic words when reviews were grouped by ratingR for Logic.

<table>
<thead>
<tr>
<th>Dim</th>
<th>Split</th>
<th>Method</th>
<th>Q1 Answer1 Mean</th>
<th>Std.Error</th>
<th>Q2 Answer2 Mean</th>
<th>Std.Error</th>
<th>Q3 Answer3 Mean</th>
<th>Std.Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flow</td>
<td>ratingR</td>
<td>Freq</td>
<td>-.217</td>
<td>.135</td>
<td>.457</td>
<td>.074</td>
<td>.217</td>
<td>.109</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TopicS</td>
<td>.413</td>
<td>.127</td>
<td>.565</td>
<td>.074</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ratingW</td>
<td>Freq</td>
<td>-.043</td>
<td>.139</td>
<td>.217</td>
<td>.074</td>
<td>.652</td>
<td>.109</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TopicS</td>
<td>.022</td>
<td>.144</td>
<td>.783</td>
<td>.061</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insight</td>
<td>ratingR</td>
<td>Freq</td>
<td>.043</td>
<td>.132</td>
<td>.522</td>
<td>.074</td>
<td>.370</td>
<td>.130</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TopicS</td>
<td>.326</td>
<td>.128</td>
<td>.543</td>
<td>.074</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ratingW</td>
<td>Freq</td>
<td>.109</td>
<td>.133</td>
<td>.283</td>
<td>.067</td>
<td>.565</td>
<td>.115</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TopicS</td>
<td>.391</td>
<td>.134</td>
<td>.717</td>
<td>.067</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logic</td>
<td>ratingR</td>
<td>Freq</td>
<td>.391</td>
<td>.118</td>
<td>.304</td>
<td>.069</td>
<td>.283</td>
<td>.138</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TopicS</td>
<td>-.174</td>
<td>.133</td>
<td>.587</td>
<td>.073</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ratingW</td>
<td>Freq</td>
<td>.109</td>
<td>.140</td>
<td>.391</td>
<td>.073</td>
<td>.261</td>
<td>.137</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TopicS</td>
<td>.152</td>
<td>.135</td>
<td>.522</td>
<td>.074</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Together</td>
<td>Freq</td>
<td>.07</td>
<td>.190</td>
<td>.36</td>
<td>.482</td>
<td>.96</td>
<td>2.068</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TopicS</td>
<td>.19</td>
<td>.923</td>
<td>.62</td>
<td>.486</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Summary of Type III F-tests significance of fixed effects of Method, Split, Dim and their interactions on all variables of all three questions. Results are presented in p-value, with significant ones highlighted with “*” (p < .05).

<table>
<thead>
<tr>
<th>Source</th>
<th>Q1 Answer1</th>
<th>Correct</th>
<th>Q2 Answer2</th>
<th>Q3 Answer3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dim</td>
<td>.907</td>
<td>.000*</td>
<td>.387</td>
<td>.289</td>
</tr>
<tr>
<td>Split</td>
<td>.000*</td>
<td>.297</td>
<td>.789</td>
<td>.055</td>
</tr>
<tr>
<td>Method</td>
<td>.196</td>
<td>.039*</td>
<td>.000*</td>
<td>na</td>
</tr>
<tr>
<td>Dim*Split</td>
<td>.015*</td>
<td>.533</td>
<td>.912</td>
<td>.226</td>
</tr>
<tr>
<td>Dim*Method</td>
<td>.008*</td>
<td>.001*</td>
<td>.364</td>
<td>na</td>
</tr>
<tr>
<td>Split*Method</td>
<td>.333</td>
<td>.863</td>
<td>.003*</td>
<td>na</td>
</tr>
<tr>
<td>Dim<em>Split</em>Method</td>
<td>.001*</td>
<td>.040*</td>
<td>.004*</td>
<td>na</td>
</tr>
</tbody>
</table>

they did express positive opinions towards the effectiveness of the topic word extraction methods (Q_listDiff), the list-layout of the topic words (Q_layout), and the usefulness of the topic-word based comparison approach for peer review analysis (Q_reviewDiff, Q_largeData and Q_approach). In addition, though participants rarely refer to the full review text (Q_textRef), they have neutral opinion towards the importance of having access to the full review text during the tasks (Q_textImp).

7.1 Task performance analysis

To further understand the impact of grouping and dimension on the utility of the topic word extraction methods, we use a mixed linear model to analyze the main effects of Split, Dim, Method as well as their interactions. Here we refer to the results of Type III F-tests as recommended in SPSS, for Type III F-tests measure the effect of the target factor in question while controlling all else in the model. A summary of the observed significant effects in our analysis is outlined in Table 4.

Can we identify the review groups by their topic words? To answer our first research question, we took a further look at the correct cases using an indicator variable “Correct” which codes correct cases as 1 and codes both incorrect and “I have no idea” as 0. When using the linear mixed model to analyze the fixed effects on Correct (as summarized in Table 4), we found that Method and Dim are significant (p < .05), while Split is not. This indicates that TopicS can generate more informative topic words than Freq, regardless of how we group the reviews, though some reviewing dimensions are naturally more difficult for capturing group properties using the topic words. We also observed significant interaction effects between Method and Dim, and among all three factors. This implies that how much better TopicS is compared to Freq is affected by how we set up the review groups for comparison (related to both Split and Dim), which corresponds to the specific investigation goals of the analysis tasks.

5Type III F-tests compute sum of square as the partial sum of squares for each effect in the linear mixed model.
Do topic words reveal patterns in writing and reviewing performance? When tested on Answer2 using the linear mixed model, only Method is found to be significant ($F(1, 530.831) = 40.015, p < .001$). An interaction exists between Method and Split ($F(1, 530.831) = 8.644, p = .003$), and among all factors ($F(1, 349.122) = 5.677, p = .004$). This tells us that using the proposed TopicS is more likely to identify review patterns that are different between groups, regardless of which dimension the reviews are on. However, the utility of topic words is also influenced by the grouping, where TopicS typically outperformed Freq when used for analyzing writing performance, especially on Flow and Insight (as shown in Table 3).

Does the proposed approach extract more informative topic words? The analysis on the fixed effects of Method above already showed that TopicS can better support users in peer review analysis. Table 3 also shows that TopicS is preferred to Freq across all tasks. And further analysis with a mixed model (Table 4) shows that such preference is not influenced by either Split or Dim.

7.2 User background analysis
With respect to user background differences, we focus on demographic factors that are related to peer-review and teaching. We investigate expPR, expSWoRD, expTA and expGW by analyzing both user satisfaction and user-study task performance.

7.2.1 Measured on user satisfaction
For each survey question, we use oneway ANOVA to examine the ratings against each background factor as a binary independent variable. Results are summarized in Table 5.

<table>
<thead>
<tr>
<th>Question</th>
<th>expPR</th>
<th>expSWoRD</th>
<th>expTA</th>
<th>expGW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q_easyness</td>
<td>3.78*</td>
<td>3.24*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q_listDiff</td>
<td></td>
<td></td>
<td>4.0*</td>
<td></td>
</tr>
<tr>
<td>Q_layout</td>
<td></td>
<td></td>
<td>4.35</td>
<td></td>
</tr>
<tr>
<td>Q_reviewDiff</td>
<td></td>
<td></td>
<td>3.83*</td>
<td>3.88</td>
</tr>
<tr>
<td>Q_largeData</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q_approach</td>
<td>1.29*</td>
<td></td>
<td></td>
<td>1.47*</td>
</tr>
<tr>
<td>Q_textRef</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q_textImp</td>
<td>2.41*</td>
<td></td>
<td></td>
<td>2.47*</td>
</tr>
</tbody>
</table>

With respect to participants’ peer-review experience, students who did peer review before (expPR = yes) generally think the review analysis tasks much easier than students who never did it before (p = .003). In particular, SWoRD users feel the proposed approach more useful than non-SWoRD users (expSWoRD = yes) in helping them identify the peer review differences (p = .014). With respect to teaching experience, it is important to note that students who have teaching experience (expTA = yes) like our idea of exploring peer reviews by comparing them in groups using their topic words (p = .039). Their feedback can somehow approximate instructors opinions towards RevExplore, which suggests the usefulness of the proposed idea for instructors to examine their peer review data in real life. While participants generally have a neutral attitude to the importance of their access to the original review in full text, students who have used SWoRD (expSWoRD = yes) or graded writing assignments (expGW = yes) before rely on this information much less than the others (p = .006, p = .048, respectively), and they think it less important than the others as well (p = .018, p = .037, respectively). This indirectly reflects the effectiveness of our topic-word approach for peer review analysis.

7.2.2 Measured on task performance
To investigate how user background factors influence the task performance, we look at all participants’ task performance across all conditions, considering expPR, expSWoRD, expTA and expGW as between-subjects effects and Method, Split and Dim as within-subjects effects. In this setting, we use the repeated-measures linear model provided by SPSS to run a Mixed Model ANOVA. First, we look for any main effect caused by the between-subjects factors; second, we examine the interactions between within- and between-subjects factors which show up in the within-subjects section of the repeated-measures analysis.

First of all, there is no significant interaction or main effect of the between-subjects factors observed on participants’ answers to any of the review analysis questions. This means that users’ prior experience in teaching and peer-review does not directly influence their task performance, which indirectly validates our using college students as the user study subjects.

However, user background factors do exert impact on the utility of topic-word analytics, as these factors qualify the effects of the within-subjects factors, especially Method, as summarized in Table 6. It is interesting to see that none of expPR, expSWoRD, expTA or expGW interacts with Method by itself alone, but in pairs. For identifying review groups (Q1), expTA occurs in both interactions (Method*expSWoRD *expTA and Method*expPR*expTA), while the other between-subjects factor is about peer review. When peering into the group differences, participants who have both teaching and peer-review experience tend to have better performance (based on modified population marginal mean). For Method*expSWoRD*expTA, SWoRD users who have TA experience exhibit better performance when using TopicS than using Freq, though such difference was not observed when we examined Mehtod*expPR*expTA. With respect to Dim (examining Dim*expPR on Q1), peer-review novels achieved their best performance on Logic, while participants who have peer-review experience did best on Insight. For both groups Flow is the most difficult dimension. Furthermore, to which extend TopicS is better than Freq is influenced by the interaction between Dim and user’s peer-review experience (expPR/expSWoRD).

In addition, we also observed that the main effects of the within-subjects factors given the presence of the between-subjects effects generally follow the pattern of Table 4 (which does not consider between-subjects effects), thus we do not discuss them here again.
In the future, we would like to evaluate the proposed approach in the context of RevExplore, which allows users to specify analytic goals at runtime. Finally we hope to integrate RevExplore into SWoRD as part of the teacher dashboard to support interactive review content analytics.

9. ACKNOWLEDGEMENTS

This work is supported by Andrew Mellon Predoctoral Fellowship 2012-2013 and IES award R305A120370.

10. REFERENCES