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. Furthermore, 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 textural data. For example, it has been used to compare political speeches from different people[1] In our study, we consider this method as a baseline (denoted as *Freq*) 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[2], 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 comments[4] 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 backwards regarding review helpfulness on a scale of 1-7, by the students who received the reviews. For our analysis, we will aggregate the helpfulness ratings for each reviewer, and use the average rating (**ratingR**) as a measure of a student's *reviewing performance*. As this step was not mandatory, the ratingR is only available for 12 students regarding their reviewing performance. (Experts' reviews were excluded in this backward evaluation.)

5. PRE-DEFINED REVIEW GROUPINGS

Different groupings of the peer reviews allow users to specify different goals in their review analysis tasks. In our user study, we look at two groupings of reviews based on existing review ratings to investigate student writing and reviewing performance. We use them as examples to examine how topic words extracted from a group of reviews can reflect the group's properties. As the instructor specified a different reviewing focus for each dimension, we consider the analysis of reviews on different dimensions as different tasks.

5.1 By paper author's average rating

To investigate student *writing performance*, we split students into high and low performance groups, based on a median split of students' **ratingW**. Then we create "high" and "low" groups of reviews accordingly, based on the group membership of the student who *received* the reviews. The hypothesis is that students who are highly rated have different writing issues compared with those who have lower ratings, and that such differences are reflected in the peer reviews that students receive.

5.2 By reviewer's average helpfulness rating

Similarly, we investigate student *reviewing performance* by splitting students into "high" and "low" groups based on a median split of their **ratingR**. Then we create the "high" and "low" review groups accordingly, based on the group membership of the student who *wrote* the reviews. We hypothesize that review topic words can reveal reviewing issues distinguishing helpful and less-helpful reviews.

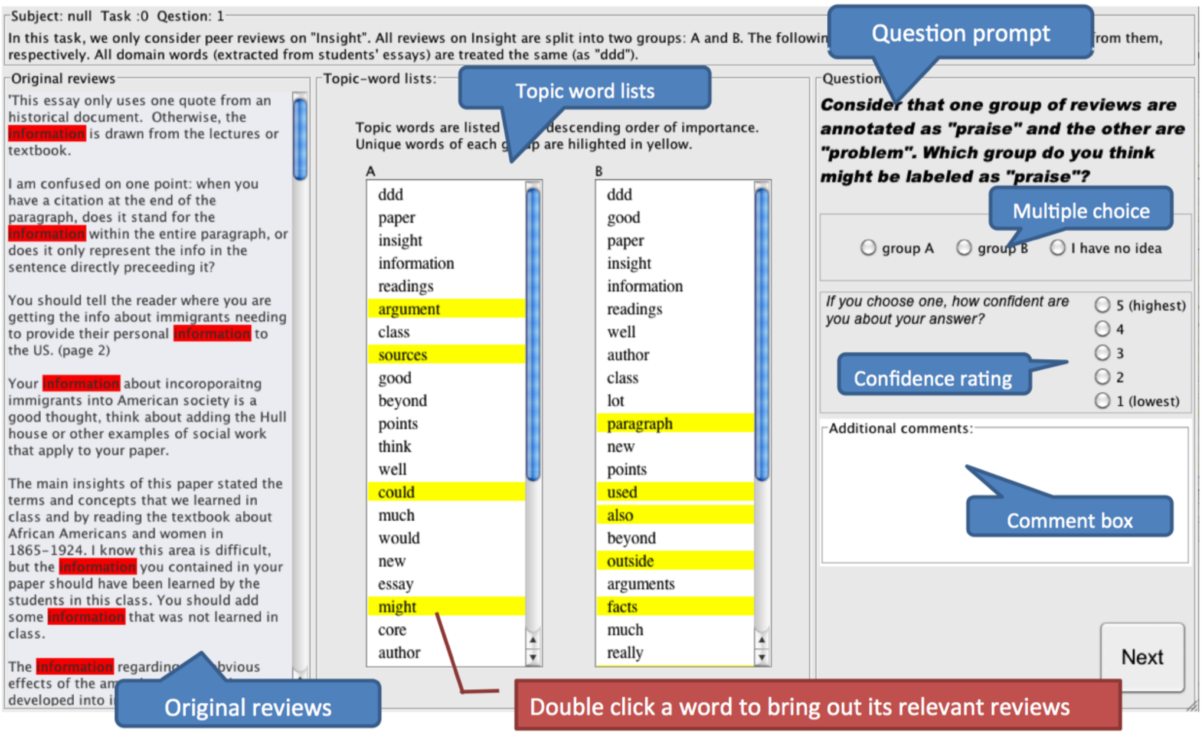
6. EXPERIMENT SETUP

We examine whether RevExplore is a useful analytic tool by evaluating the topic-word analytic approach in the context of educational peer-review analysis. In particular, we compare the effectiveness of **two** topic-word extraction methods (*TopicS* and *Freq*) quantitatively using six real peer-review analysis tasks – **two** review groupings (ratingW and ratingR) across **three** reviewing dimensions (Flow, Logic, and Insight). We denote the three factors as *Method*, *Split* and *Dim* respectively. We conduct a formative user study using a $2 \times 2 \times 3$ within-subject design, in which every subject goes through all experimental conditions in random order. In this paper, we analyze users' task performance and user satisfaction both qualitatively and quantitatively.

For task performance analysis, we are interested in three research questions: 1) whether humans can identify the review groups based on their topic words; 2) whether it is feasible to identify any pattern of student peer review performance by comparing peer reviews in groups based on their topic words; 3) whether topic words learned by *TopicS* are more informative than those by *Freq*. Thus we accordingly asked

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

[4] A single peer review can have multiple comments.



Subject: null Task :0 Qestion: 1

In this task, we only consider peer reviews on "Insight". All reviews on Insight are split into two groups: A and B. The followin from them, respectively. All domain words (extracted from students' essays) are treated the same (as "ddd").

Original reviews

'This essay only uses one quote from an historical document. Otherwise, the information is drawn from the lectures or textbook.

I am confused on one point: when you have a citation at the end of the paragraph, does it stand for the information within the entire paragraph, or does it only represent the info in the sentence directly preceeding it?

You should tell the reader where you are getting the info about immigrants needing to provide their personal information to the US. (page 2)

Your information about incoroporaitng immigrants into American society is a good thought, think about adding the Hull house or other examples of social work that apply to your paper.

The main insights of this paper stated the terms and concepts that we learned in class and by reading the textbook about African Americans and women in 1865-1924. I know this area is difficult, but the information you contained in your paper should have been learned by the students in this class. You should add some information that was not learned in class.

The information regarding bvious effects of the am developed into i

Topic-word lists:

Topic words are listed descending order of importance. Unique words of each g p are hilighted in yellow.

A	B
ddd	ddd
paper	good
insight	paper
information	insight
readings	information
argument	readings
class	well
sources	author
good	class
beyond	lot
points	paragraph
think	new
well	points
could	used
much	also
would	beyond
new	outside
essay	arguments
might	facts
core	much
author	really

Question

Consider that one group of reviews are annotated as "praise" and the other are "problem". Which group do you think might be labeled as "praise"?

○ group A ○ group B ○ I have no idea

If you choose one, how confident are you about your answer?

○ 5 (highest)
○ 4
○ 3
○ 2
○ 1 (lowest)

Additional comments:

Figure 1: Interface annotation for peer review analysis user study.

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 significant 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.

Table 1: User distribution over demographic factors that are related to peer-review and teaching.

Factor	Frequency_no	Frequency_yes
$expPR$	23	23
$expSWoRD$	29	17
$expTA$	23	23
$expGW$	29	17

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

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 ($p < 0.05$).

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

questions. During the study, if participants feel that some topic word is hard to interpret, they can double click the word on the list to bring out its related reviews in the original review pane. The overall length of the user study was about an hour.

During the user study, the participants completed all tasks in random order. For each task, we computed the same number of topic words for the high and low review groups using *Freq* and *TopicS*, and randomly picked one extraction method for a participant to examine first. For a given method, we presented the corresponding topic words in two lists, one for each group. And we asked participants the same questions Q1 and Q2 regarding the group differences without revealing the group labels. In order to exclude the impact of revealing the group labels for examining the first method, when switching to the second extraction method, we randomly layout the two list of topic words computed by the second method and then asked Q1 and Q2 again. After participants visited both methods, we allowed them to revisit the topic words computed by both methods with correct group labels attached, asking them to vote on which method generated more informative words in terms of identifying the different review focus between the two groups (Q3). In Q1, participants needed to provide their prediction or check "I have no idea"; in Q2, participants needed to answer either yes or "no", and they could also articulate what patterns they found in free text; in Q3, participants needed to vote for the better method or check "no preference".

After the user study, the participants took an exit survey to rate the utility of the two methods as well as the topic-word analytics in general for analyzing students' peer reviews. There are eight subjective questions in the survey. Participants gave their opinions in a scale of 5 points, with 3 being neutral. Survey questions and the descriptive statistics of user satisfaction are presented in Table 2.

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:

$$Answer1 = \begin{cases} 1 & \text{if the answer is correct,} \\ -1 & \text{if the answer is incorrect,} \\ 0 & \text{if ``I have no idea''.} \end{cases} \tag{1}$$

$$Answer2 = \begin{cases} 1 & \text{if yes,} \\ 0 & \text{if ``no''.} \end{cases} \tag{2}$$

$$Answer3 = \begin{cases} 1 & \text{if vote for } \textit{TopicS,} \\ -1 & \text{if vote for } \textit{Freq,} \\ 0 & \text{if ``no preference''.} \end{cases} \tag{3}$$

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.

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

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 "*" ($p < .05$).

	Q1		Q2	Q3
Source	*Answer1*	*Correct*	*Answer2*	*Answer3*
Dim	.907	.000*	.387	.289
Split	.000*	.297	.789	.055
Method	.196	.039*	.000*	na
Dim*Split	.015*	.533	.912	.226
Dim*Method	.008*	.001*	.364	na
Split*Method	.333	.863	.003*	na
Dim*Split*Method	.001*	.040*	.004*	na

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[5] 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.

[5] Type 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 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 5: Oneway ANOVA analysis of user-background factors (binary) on user satisfaction. Factors that are significant ($p < 0.05$, highlighted with "*") or in trend are denoted by the mean value of the "yes" group.

Question	expPR	expSWoRD	expTA	expGW
Q_easyness	3.78*	3.24*		
Q_listDiff				
Q_layout				
Q_reviewDiff		4.0*		
Q_largeData		4.35		
Q_approach			3.83*	3.88
Q_textRef		1.29*		1.47*
Q_textImp		2.41*		2.47*

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 ($p = .033$). 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 $Mehtod{*}expPR{*}expTA$), 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.

Table 6: Summary of Mixed Model ANOVA of within-subjects effects, including interactions between user-background factors (between-subjects effects) and Method, Split, Dim (within-subjects effects). Significant results are presented in p-value ($p \leq .05$).

	Q1		Q2	Q3
Source	*Answer1*	*Correct*	*Answer2*	*Answer3*
Dim*expPR	$F(2,66) = 3.1$, $p = .050$	$F(2,66) = 3.6$, $p = .032$		
Method*expSWoRD*expTA	$F(1,33) = 7.4$, $p = .001$	$F(1,33) = 6.1$, $p = .019$		na
Mehtod*expPR*expTA	$F(1,33) = 9.6$, $p = .004$	$F(1,33) = 4.2$, $p = .049$		na
Mehtod*expTA*expGW			$F(1,33) = 6.1$, $p = .019$	na
Dim*Method*expPR	$F(1,66) = 3.4$, $p = .040$			na
Dim*Method*SWoRD	$F(2,66) = 4.0$, $P = .022$	$F(2,66) = 5.5$, $p = .006$		na

8. CONCLUSIONS

In this paper we evaluate the topic-word analytics for analyzing educational peer reviews with a user study. The user study shows that student peer reviews can be used to examine student writing and reviewing performance based on peer review topic words, and that the proposed comparison-oriented topic-word extraction method (*TopicS*) suits our analytic tasks best compared with the frequency based method (*Freq*). However, the utility of the learned topic words is influenced by the analytic goals (specified through review grouping) and dimensions, as well as users' prior experience in teaching and peer-review. Analysis of user satisfaction shows that participants who have teaching experience significantly favor our approach more than the others, which suggests the usefulness of the proposed approach in supporting instructors for analyzing student peer reviews in the real-world. Even though we did not include manual digestion of original peer reviews as a baseline, we indirectly compare it with our topic-word approach in the exit survey (Q_approach).

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

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] J. Boyd-Graber, J. Chang, S. Gerrish, C. Wang, and D. Blei. Reading tea leaves: How humans interpret topic models. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, 2009.

[3] K. Cho. Machine classification of peer comments in physics. In *Proceedings of the First International Conference on Educational Data Mining*, pages 192–196, 2008.

[4] K. Cho and C. D. Schunn. Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers and Education*, 48(3):409–426, 2007.

[5] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 19(1):61–74, 1993.

[6] I. Goldin and K. Ashley. Peering inside peer review with bayesian models. In *Artificial Intelligence in Education*, pages 90–97. Springer, 2011.

[7] B. Lehman, W. L. Cade, and A. Olney. Off topic conversation in expert tutoring: Waste of time or learning opportunity. In *EDM'10*, pages 101–110, 2010.

[8] C.-Y. Lin and E. Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics*, volume 1 of *COLING '00*, pages 495–501, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[9] F.-R. Lin, L.-S. Hsieh, and F.-T. Chuang. Discovering genres of online discussion threads via text mining. *Computers & Education*, 52(2):481–495, 2009.

[10] N. Ming and E. Baumer. Using text mining to characterize online discussion facilitation. *Journal of Asynchronous Learning Networks*, 2011.

[11] M. M. Nelson and C. D. Schunn. The nature of feedback: how different types of peer feedback affect writing performance. *Instructional Science*, 37:375–401, 2009.

[12] A. Nenkova and A. Louis. Can you summarize this? identifying correlates of input difficulty for generic multi-document summarization. In *Proceedings of Association for Computational Linguistics*, 2008.

[13] W. Xiong and D. Litman. Automatically predicting peer-review helpfulness. In *Proceedings 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

[14] W. Xiong, D. Litman, J. Wang, and C. Schunn. An interactive analytic tool for peer-review exploration. In *Proceedings of the 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, pages 174–179, 2012.

[15] W. Xiong and D. J. Litman. Identifying problem localization in peer-review feedback. In *Proceedings of Tenth International Conference on Intelligent Tutoring Systems*, volume 6095, pages 429–431, 2010.

[16] Z. Zhai, B. Liu, H. Xu, and P. Jia. Constrained lda for grouping product features in opinion mining. *Advances in Knowledge Discovery and Data Mining*, pages 448–459, 2011.