A User Study on the Automated Assessment of Reviews

Lakshmi Ramachandran and Edward F. Gehringer
North Carolina State University
{lramach,efg}@ncsu.edu

ABSTRACT

Reviews are text-based feedback provided by a reviewer to the author of a submission. Reviews play a crucial role in providing feedback to people who make assessment decisions (e.g. deciding a student's grade, purchase decision of a product). It is therefore important to ensure that reviews are of a good quality. In our work we focus on the study of academic reviews. A review is considered to be of a good quality if it can help the author identify mistakes in their work, and help them learn possible ways of fixing them. *Metareviewing* is the process of evaluating reviews. An automated metareviewing process could provide quick and reliable feedback to reviewers on their assessment of authors' submissions. Timely feedback on reviews could help reviewers correct their assessments and provide more useful and effective feedback to authors. In this paper we investigate the usefulness of metrics such as *review relevance, content type, tone, quantity* and *plagiarism* in determining the quality of reviews. We conducted a study on 24 participants, who used the automated assessment feature on Expertiza, a collaborative peer-reviewing system. The aim of the study is to identify reviewers' perception of the usefulness of the automated assessment feature and its different metrics. Results suggest that participants find relevance to be the most important and quantity to be the least important in determining a review's quality. Participants also found the system's feedback from metrics such as content type and plagiarism to be most useful and informative.

Keywords

review quality assessment, metareview metrics, user experience survey

1. INTRODUCTION

In recent years there has been a considerable amount of research directed towards developing educational systems that foster collaborative learning. Collaborative learning systems provide an environment for students to interact with other students, exchange ideas, provide feedback and use the feedback to improve their own work. Systems such as SWoRD [1] and Expertiza [3] are web-based collaborative peer-review systems, which promote team work by allowing students to build shared knowledge with an exchange of ideas. These systems also provide an environment for students to give feedback to peers on their work.

The process of providing feedback to peers on their work may help students learn more about the subject, and develop their critical thinking. Rada et al. found that students who evaluated their peers' work were more likely to improve the quality of their own work than those students who did not provide peer reviews [4]. The peer review process may also help students learn to be more responsible.

The classroom peer review process is very much similar to the process of reviewing scientific articles for journals. Scientific reviewers tend to have prior reviewing experience and a considerable knowledge in the area of the author's submission (the text under review). Students on the other hand are less likely to have had any prior reviewing experience. They have to be guided to provide good quality reviews that may be useful to their peers.

Metareviewing can be defined as the process of reviewing reviews, i.e., the process of identifying the quality of reviews. Metareviewing is a manual process [2, 5, 6] and just as with any process that is manual; metareviewing too is (a) slow, (b) prone to errors and (c) likely to be inconsistent - the set of problems, which makes automated metareviewing necessary. An automated metareview process ensures consistent, bias-free reviews to all reviewers. This also ensures provision of immediate feedback to reviewers, which is likely to motivate them to improve their work and provide more useful feedback to the authors.

In this work we propose the use of a system that automatically evaluates student review responses. We use a specific set of metrics such as *review's relevance* to the work under review (or the submission), the *type of content* a review contains, *tone* of the review, *quantity* of feedback provided and presence of *plagiarism*, to carry out metareviewing. We have integrated the automated metareview feature (with the listed set of metrics) into Expertiza [3]. Expertiza is a collaborative web-based learning application. A screenshot of the metareview output from the system is shown in Figure1. We have conducted an exploratory analysis to study the importance of the review quality metrics and usefulness of the system's outputs, as judged by users of the system.

2. RELATED WORK

One of the earlier approaches to manually assessing the quality of peer reviews involved the creation and use of a Review Quality Instrument (RQI) [9]. Van Rooyen et al. use the RQI to check whether a reviewer discusses the following - (1) importance of the

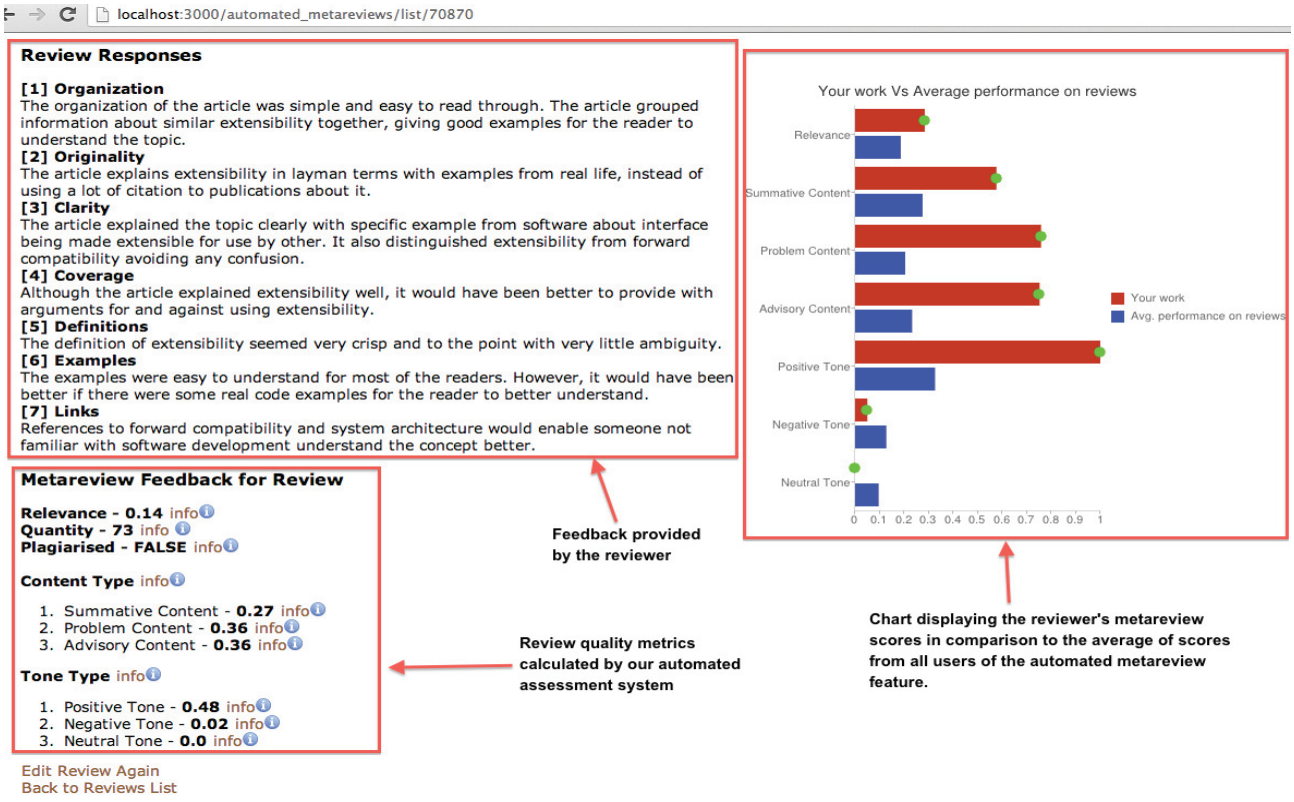

Figure 1: Output from the automated metareview feature on Expertiza [3]. We provide a comparison of the participant reviewer's scores with other reviewers' metareview scores (in a chart) to help reviewers gauge how well they are doing on a certain metric.

research question, (2) originality, (3) strengths and weaknesses, (4) presentation and interpretation of results. In addition, the RQI also checks if a review was constructive, and if the reviewer had substantiated his/her claims. We incorporate some of these metrics in our approach, e.g. detecting constructiveness in reviews (based on its content), checking whether reviewers substantiated their claims (by identifying relevance to the author's submission), to automatically assess review quality.

Nelson and Schunn studied feedback features that help authors understand and use reviews [10]. They found that features such as problem localization and solution suggestion helped authors understand feedback. These are some of the types of content we look for during review content identification.

Kuhne et al. use authors' ratings of reviews to identify the quality of peer reviews [5]. They found that authors are contented with reviewers who appear to have made an effort to understand their work. This finding is useful to our automatic review quality assessment system, which assesses reviews based on the usefulness of its content. Our system also detects the relevance of reviews, which may be indicative of the effort made by a reviewer to understand the author's work and provide specific feedback.

Xiong et al. look for problems identified by reviewers in the author's work in peer reviews from the SWoRD system [11]. Xiong et al. use a bag-of-words, exact match approach to detect problem localization features. They use a shallow semantic match approach, which uses counts of nouns, verbs etc. in the text as features. Their approach does not incorporate relevance identification nor does it identify content type. Cho uses machine classification techniques such as naïve Bayes, support vector machines (SVM) and decision trees to classify review comments [12]. Cho manually breaks down every peer comment into idea units, which are then coded as a praise, criticism, problem detection, solution suggestion, summary or off-task comment.

Some other approaches used to study the usefulness of reviews are those by Turney [15], Dalvi [16] and Titov [17]. Peter D. Turney uses semantic orientation (positive or negative) to determine whether a review can be classified as recommended or not recommended. Turney's approach to differentiate positive from negative reviews involves identifying similarity between phrases containing adverbs and adjectives to terms "excellent" and "poor". Turney uses semantic orientation to recommend products or movies. We also use semantic orientation (referred to as tone) to identify the degree of sensitivity with which reviewers convey their criticisms.

Lim et al. identify reviewers who target e-commerce products and applications, and generate spam reviews [18]. The problem of spamming may be analogous to the problem of copy-pasting text in order to game the automated assessment system into giving reviewers high scores on their reviews. Therefore, we use a metric to detect plagiarized reviews.

There exist research works that discuss metrics that are important in review quality identification, and some that apply shallow approaches to determine quality. However, there is no work that takes factors such as relevance, content type, tone, quantity and plagiarism into consideration while determining review quality. Our sys-

Table 1: Some examples of reviews.

S No.	Review
1	"The example needs work."
2	"Yes, the organization is poor."

tem is an amalgamation of existing research in the said areas. In the next section, we provide an overview of the different review quality metrics.

3. AUTOMATED REVIEW ASSESSMENT

In order to assess quality, reviews have to be represented using metrics that capture their most important features. In general a good review contains: (1) coherent and well-formed sentences, which can be easily comprehended by the author, and (2) sufficient amount of feedback. In this section we discuss the metrics we use to assess reviews.

3.1 Review relevance

Reviewers may provide vague, unjustified comments. Comments in Table 1 are generic, and do not refer to a specific object in the text under review. For instance, what type of "work" does the "example" need or, what is poor about the "organization"? These reviews are ambiguous, and need to be supported with more information. Also, how do we know if the review has been written for the right submission, for instance any article may contain an example. Review relevance helps identify if a review is talking about the right submission.

We identify relevance in terms of the semantic and syntactic similarities between two texts. We use a word order graph, whose vertices, edges and double edges help determine structure-based match across texts. We use WordNet to determine semantic relatedness. Our approach has been described in Ramachandran and Gehringer [19].

3.2 Review content

A review is expected to provide an assessment of the kind of work that was done - praising the submission's positive points, identifying problems, if any, and offering suggestions on ways of improving the submission. Review examples in Table 1 do not provide any details. Reviews must identify problems in the author's work, and provide suggestions for improvement in order to be useful to authors, thus helping them understand where their work is lacking or how it can be improved. Content of a review identifies the type of feedback provided by the reviewer. We look for the following types of content in a review:

- **Summative** - Summative reviews contain either a positive or a neutral assessment of the author's submission. *Example:* "I guess a good study has been done on the tools as the content looks very good in terms of understanding and also originality. Posting reads well and appears to be largely original with appropriate citation of other sources."
- **Problem detection** - Reviews in this category are critical of the author's submission and point out problems in the submission. *Example:* "There are few references used and there are sections of text quoted that appear to come from a multitude of web sites." However, problem detection reviews only find problematic instances in the author's work, and do not offer any suggestions to improve the work.
- **Advisory** - Reviews that offer the author suggestions on ways of improving their work fall into this category. *Example:* "Although the article makes use of inline citations which is a plus, there are only a few references. Additional references could help support the content and potentially provide the examples needed." Advisory reviews display an understanding of the author's work, since the reviewer has taken the effort to provide the author with constructive feedback.

Different types of review content have different degrees of usefulness. For instance summative reviews provide only summaries of the author's work and are less useful to the author, whereas reviews that identify problems in the author's work or provide suggestions can be used by authors to improve their work, and are hence considered more important. We use a cohesion-based pattern identification technique to capture the meaning of a class of reviews.

3.3 Review tone

Tone refers to the semantic orientation of a text. Semantic orientation depends on the reviewer's choice of words and the presentation of a review. Tone of a review is important because while providing negative criticism reviewers might unknowingly use words that may offend the authors. Therefore we use tone information to help guide reviewers. A review can have one of three types of tones - positive, negative or neutral. We look for positively or negatively oriented words to identify the tone of a review [15]. We use positive and negative indicators from an opinion lexicon provided by Liu et al [20] to determine the semantic orientation of a review. Semantic orientation or tone of the text can be classified as follows:

- **Positive** - A review is said to have a positive tone if it predominantly contains positive feedback, i.e., it uses words or phrases that have a positive semantic orientation. *Example:* "The page is very well-organized, and the information is complete and accurate." Adjectives such as "well-organized", "complete" and "accurate" are good indicators of a positive semantic orientation.
- **Negative** - This category contains reviews that predominantly contain words or phrases that have a negative semantic orientation. Reviews that provide negative criticism to the author's work fall under this category, since while providing negative remarks reviewers tend to use language that is likely to offend the authors. Such reviews could be morphed or written in a way that is less offensive to the author of a submission. *Example:* "The examples are not easy to understand and have been copied from other sources. Although the topic is Design Patterns in Ruby, no examples in Ruby have been provided for Singleton and Adapter Pattern."

  The given example contains negatively oriented words or phrases such as "not easy to understand" ,"copied", "no examples". Review segment "..have been copied from other sources..." implies that the author has plagiarized, and could be construed by the author as a rude accusation. One of the ways in which this review could be re-phrased to convey the message, so as to get the author to acknowledge the mistake and make amends, is as follows - "The topic on Design Patterns in Ruby could be better understood with more examples, especially for the Singleton and Adapter patterns. Please try to provide original examples from your experience or from what was discussed in class."

- **Neutral** - Reviews that do not contain either positively or negatively oriented words or phrases, or contain a mixture of both are classified into this category. *Example:* "The organization looks good overall. But lots of IDEs are mentioned in the first part and only a few of them are compared with each other. I did not understand the reason for that." This review contains both positively and negatively oriented segments, i.e., "The organization looks good overall" is positively oriented, while "I did not understand the reason for that." is negatively oriented. The positive and negatively oriented words when taken together give this review a neutral orientation.

In case of both content and tone, a single review may belong to multiple categories. For instance consider the review, "Examples provided are good; a few other block structured languages could have been talked about with some examples as that would have been pretty useful to give a broader pool of languages that are block structured." While classifying for content, we see that the first part of the review, "Examples provided are good" praises the submission, while the remaining part of the review provides advice to the author. Our content identification technique identifies the amount of each type of content or tone (on a scale of 0 - 1) a review contains. Similarly in the case of tone, we identify the degree of positive, negative or neutral orientation of each review.

3.4 Review quantity

Text quantity is important in determining review quality since a good review provides the author with sufficient feedback. We plan on using this metric to indicate to the reviewer the amount of feedback they have provided in comparison to the average review quantity (from other reviewers of the system), thus motivating reviewers to provide more feedback to the authors. We identify quantity by taking a count of all the unique tokens in a piece of review. For instance, consider the following review, "The article clearly describes its intentions. I felt that section 3 could have been elaborated a little more." The number of unique tokens in this review is 15 (excluding articles and pronouns).

3.5 Plagiarism

Reviewers tend to refer to content in the author's submission in their reviews. Content taken from the author's submission or from some external source (Internet) should be placed within quotes in the review. If reviewers copy text from the author's submission and fail to place it within quotes (knowingly or unknowingly) it is considered as *plagiarism*.

Each of the review quality metrics listed is determined independently, and are integrated into a complete review quality assessment system. Reviewers are given feedback on each of these listed metrics, so that they get a complete picture of the completeness and quality of their review.

4. USER EXPERIENCE STUDY

We decided to study the experience of using an automated metareview system, since different types of reviewers - students, teaching assistants and faculty may use this feature. We study the extent to which users of an automated quality assessment system would perceive it to be important, and the output of the system to be useful. The study is important because it helps us understand whether reviewers learn and benefit from such an automated metareview system. This study also helps us learn what aspects of the feature can be improved, by identifying what the surveyed reviewers liked or disliked about the feature. A positive experience from using this feature may mean that reviewers would be more inclined to use it to improve their reviews.

According to Kuniavsky [21], user experience is "the totality of end-users' perceptions as they interact with a product or service. These perceptions include effectiveness (how good is the result?), efficiency (how fast or cheap is it?), emotional satisfaction (how good does it feel?), and the quality of the relationship with the entity that created the product or service (what expectations does it create for subsequent interactions?)." There exist several other definitions for the term *user experience* (abbreviated as UX) [22]. UXMatters[1] defines user experience as that which "Encompasses all aspects of a digital product that users experience directly - and perceive, learn, and use - including its form, behavior, and content." They also state that "Learnability, usability, usefulness, and aesthetic appeal are key factors in users' experience of a product." Therefore, apart from a study of factors such as user's perceptions, feelings or responses to a system, a user experience survey should also involve a study of the learning gained from a system and the usefulness of a system.

The aim of this study is to identify the degree of importance participants attach to each of the metareview metrics–review relevance, content, tone, quantity and plagiarism. This study will help us identify how effective the system is at helping reviewers learn about characteristics of their reviews.

5. EXPERIMENTS

To study the usefulness of our review quality assessment system we investigate the following broad research questions:

RQ1: *Do automated metareviews provide useful feedback?*
RQ2: *Which of the review quality metrics are more or less important than the others?*
RQ3: *Which of the review quality metrics' output did the reviewers find more or less useful when compared to the others?*

5.1 Participants

In order to identify how useful users of the automated metareview feature find it to be, we recruited 24 participants to (1) use the feature on Expertiza and (2) provide us with information on their experience by filling out a survey. Participants were recruited with an email message, which explained to them the purpose of the study. The set of participants included 15 doctoral students, 3 masters' students and 1 undergraduate student, all of whom were from the computer science department at North Carolina State University, and 5 research scientists from academia and industry.

5.2 Data collection

Our data collection process involved two steps. In the first step, participants were asked to use the automated metareview feature on Expertiza. They use the system to write a review for an article. For our study, we chose a wiki article on *Software Extensibility*[2]. We chose this article since we were recruiting subjects from the field of computer science, and Software Extensibility is a topic most computer science students or researchers are familiar with. A detailed

[1]UXMatters - User experience definition - http://www.uxmatters.com/glossary/

[2]Software Extensibility - `https://en.wikipedia.org/wiki/Extensibility`

Table 2: Detailed set of instructions to help complete the survey

1.	Use username/password to log into Expertiza.
2.	Click on assignment "User Study"
3.	Click on "Others' Work" (Since you will be reviewing someone else's work.)
4.	Click on "Begin" to start the review.
5.	Click the url under the "Hyperlinks" section. Read the article on Software Extensibility. Please keep in mind that you are reviewing this article.
6.	Answer questions on the review rubric describing the quality of the article you read. After answering all the review questions, click on the "Save Review" button.
7.	Wait for a few minutes for the system to generate the automated feedback.
8.	Fill out the **user-experience questionnaire**.

set of instructions was provided to each of the participants to help them complete the study (Table 2).

A review rubric is provided to the participants to help them write the review. The rubric contains questions on the organization, originality, clarity and coverage of the article under review. The rubric also evokes information on quality of the definitions, examples and links found in the article.

When participants submit their reviews, they are presented with automated feedback from our system. This feedback gives them information on different aspects of their review such as (1) content type, (2) relevance of the review to the article, (3) tone, (4) quantity of text and (5) presence of plagiarism. A screenshot of the output is available in Figure1. The participant reviewer reads and understands the metareview feedback.

In the second step of data collection, the participant reviewer is asked to fill out a user experience questionnaire (Step 8 in Table 2). The user experience questionnaire is a big part of this study, and has been explained in detail in Section 6.

6. USER EXPERIENCE QUESTIONNAIRE

The user experience questionnaire consists of four sections - *participant background, importance of reviews, importance of metrics, usefulness of system's output*. The questions we use in our user experience survey are discussed in the following sections. Answers to each of these questions are given on a scale of 1 (lowest) to 5 (highest).

6.1 Participant background

In the *background* section, participants were questioned about their experience in writing reviews, and in their experience with using peer-review systems such as Expertiza. The exact questions were:

Q1: *Do you have prior reviewing experience?*
Q2: *Do you have prior experience using the Expertiza system?*
Q3: *Have you used a peer-review system before?*
Q4: *Are you a(n): Undergraduate, Masters or PhD student, or Other?*

6.2 Importance of reviews and metareviews

In the *importance* section, we questioned participants on the importance of reviews and metareviews to a system.

Q5: *How important do you think reviews are in a decision-making process?*
Q6: *How important do you think metareviews (review of a review) are in a decision-making process?*

Answers are given on a 5-point scale - *unimportant, somewhat important, neutral, important* and *extremely important*. This section also includes an open question to gather textual feedback from participants. All these questions are optional, i.e., the participant may choose not to respond to any of them.

We also gauge whether participants would be motivated to use reviews to improve the quality of their submission (as an author), and metareviews to improve the quality of their reviews (as a reviewer). We therefore included the following questions in the questionnaire:
Q7: *Would better reviews inspire you to use the feedback in your revisions?*
Q8: *Would automated metareviews motivate you to update your reviews?*
Q9: *Do the automated metareviews provide useful feedback?*

6.3 Importance of metareview metrics

In the *importance of metrics* section we identify how important participants think the different metareview metrics are in gauging the quality of a review.

Q10: *How important do you think each of the review quality metrics is in learning about the quality of your review? 1. Review relevance, 2. Review content 3. Tone 4. Quantity 5. Plagiarism*

The answers are given on a 5-point scale. This question helps us identify the metrics to which users of the system attach most importance, or to which ones they attach the least importance. This section also allows participants to provide any additional comments, to learn about the participants' opinions of the different metrics, or any other related information.

6.4 Usefulness of system's metareview output

This section helps us study the usefulness of the system's outputs. These questions gauge whether reviewers learned something about their review's quality from the automated feedback.

Q11: *How useful do you think the output from each of the review quality metrics is (from what you saw on Expertiza)? 1. Relevance, 2. Review content 3. Tone 4. Quantity 5. Plagiarism*

Answers are given on a 5-point scale and range from *not useful, somewhat useful, neutral, useful* or extremely useful. The ratings indicate usefulness of the chosen design for the system' output. These questions help us learn whether participants are able to successfully comprehend the meaning of the system's output. This information coupled with the information from the previous question on *importance of metrics* would help us identify the set of metrics that need improving. This section also includes an open question to gather any other comments participants may have on the system's output.

6.5 Other metrics

We included an open question on the survey to learn about any other review quality metrics, which participants think would be useful in an automated metareview system.

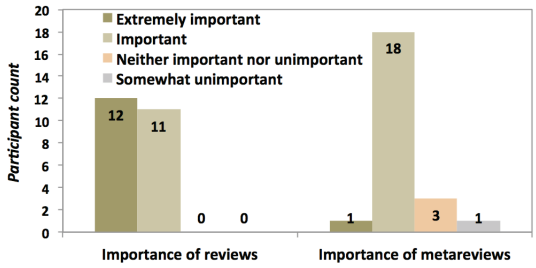

Figure 2: Participants' rating of importance of reviews and metareviews.

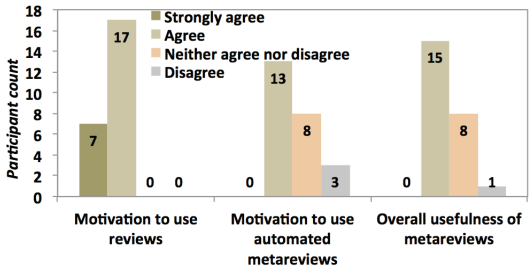

Figure 3: Participants' rating of motivation to use reviews and metareviews to improve the quality of their submission or review respectively. The chart also contains participants' estimation of usefulness of the automated metareview feature's output.

Q12: *What other information do you think might help you improve your review quality? Are there any specific review features you would like to get feedback on? e.g. language of the review, grammar, vocabulary, or nothing else*

The next section discusses our analyses on the collected data.

7. ANALYSIS OF DATA

In this section we discuss some of the findings from our data. Out of the 24 participants, 19 had prior reviewing experience. Only 7 of the participants had prior experience with the Expertiza system.

7.1 Importance of reviews and metareviews

All of the participants agreed that reviews play an important role in the decision-making process (Figure 2). A majority of the participants also agreed on the importance of metareviews (review of reviews). One participant did not respond to these questions.

We asked participants whether good quality reviews would motivate them to fix their submission. All participants agreed (7 agreed strongly) that they would incorporate suggestions from the feedback in their work (Figure 3). We asked participants whether automated feedback on their reviews would inspire them to improve their reviews. Out of the 24 participants 13 agreed that they would use the automated feedback. However 8 participants displayed doubt in the use of automated metareview feedback by answering *neither agree nor disagree*. A small number said that they would not be inclined to use the automated metareview feedback to improve their reviews.

Thus we see that as authors, participants agree that good quality feedback would motivate them to fix their work, but as reviewers they may not be inclined to use metareview feedback to update their reviews (and help other authors improve their work). The concept of automated assessment of reviews is new, and a lack of understanding of the purpose of these metrics could be one of the reasons why reviewers felt that automated metareviews may not motivate them to fix their reviews.

7.2 Importance of the review quality metrics

We analyze how participants judge each of the automated metrics' importance. The results are displayed in Figure 4. The metric, which participants rated as the most important is *relevance*. Out of the 24 participants 23 agree that relevance is important in assessing the quality of a review (3 thought it was extremely important). The next most important metric was found to be *review content*, with

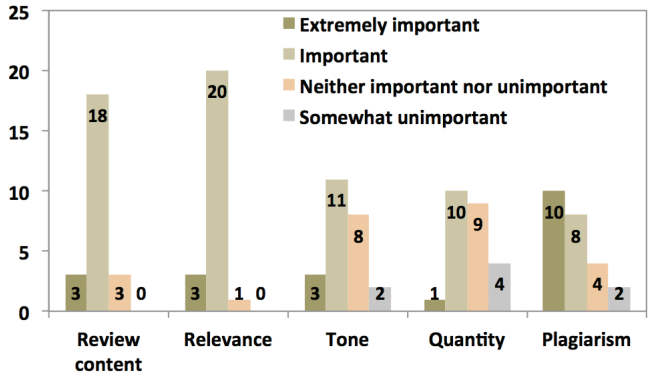

Figure 4: Participants' rating of the importance of each review quality metric.

21 of the participants agreeing on its importance (3 thought it was extremely important).

Participants found quantity to be the least important metric, with 9 of them expressing doubts on its usefulness (neither important nor unimportant) and 4 of them describing it as somewhat unimportant. Wilcoxon rank-sum test is used to determine if two metrics' ratings have identical distributions (null hypothesis) [23]. We use this test to compare metric quantity with metrics relevance and content (which have been identified as the most important metrics) at 0.05 significance level. The p value for the test on metrics quantity and relevance is 0.0003, and for metrics quantity and content is 0.002. Since these p values are < 0.05, we conclude that quantity's ratings are significantly different from those of the most important metrics - relevance and content.

Quantity contains the number of unique tokens in a review text, and is meant to motivate reviewers to write more feedback. Quantity may be obvious to a reviewer, since they are aware of the amount of feedback they have provided. Hence quantity may turn out to be the least effective, when compared with the other metrics, in conveying any new information to the reviewer. This could be why quantity is ranked as the least important quality metric.

7.3 Usefulness of system output

We questioned participants on the usefulness of the system's metareview output, to study how informative or understandable they find

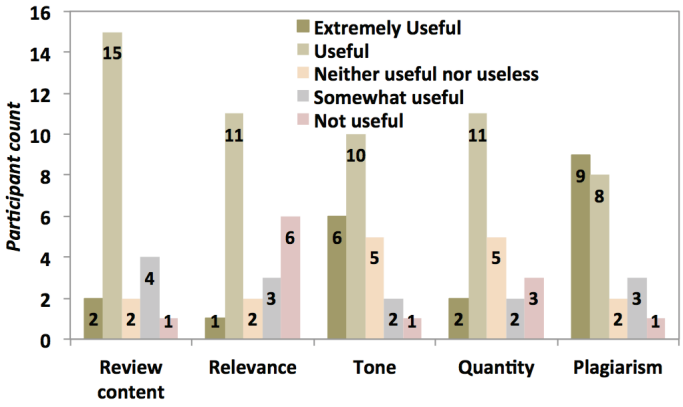

Figure 5: Participants' rating of the usefulness of each review quality metric.

it. The results of studying usefulness of metrics are displayed in Figure 5. The metrics participants rated as most useful are *plagiarism* and *review content*, with 17 of participants (9 found plagiarism extremely useful, and 2 found content extremely useful) agreeing that these metrics were useful in helping them understand where their reviews are lacking.

Tone is the second most useful metric with 16 of the participants agreeing on its usefulness, despite having 8 participants judging it to be neither important nor unimportant (from previous section). Similarly in the case of quantity, 13 of the participants found the systems' output for quantity to be useful (2 of them thought it was extremely useful), although 9 of the participants said that they thought it to be neither important nor unimportant (Figure 4).

We use the Wilcoxon test (at a significance level of 0.05) to determine if there is a significant difference (increase) in the distribution of the importance and usefulness ratings of quantity. We selected pairs, whose ratings for usefulness showed an increase from their corresponding importance ratings. The ratings have a p value of $0.03 < 0.05$, which indicates that the increase in usefulness ratings is significant. Similarly, when identifying the significance of increase between the importance and usefulness ratings of tone, we get a p value of 0.09. Although this is not < 0.05, we see that the low p value may be indicative that the improvement in usefulness ratings is not a chance occurrence (i.e., it is significant). Thus we see that although participants thought initially that tone and quantity may not be important to a metareview assessment system, they found the output from the system for these two metrics to be insightful.

Despite being judged as the most important review assessment metric only 12 of the participants found the output of the relevance metric to be useful. One of the participants expressed difficulty in interpreting the meaning of the relevance score. Our metareview feedback contains only real-valued scores in the range 0 - 1, which may not have been very useful to the reviewer in understanding the degree of relevance. This could have caused the relevance's usefulness ratings to be lower when compared to the ratings of metrics such as plagiarism, which contains true/false as output.

In the future we are planning to improve the format of the output by providing textual feedback in addition to the numeric feedback. The feedback will point to specific instances of a review that need improvement. This may make it easy for reviewers to interpret the numeric score, and maybe further motivate reviewers to use the information to improve their reviews.

7.4 Other metrics

Some of the other metrics that participants exclaimed their interest in are the *grammar and syntax* of reviews. One of the participants suggested the use of *sentence structure variability* across sentences as a means of assessing a review. The participant suggested that though short phrases may succeed in communicating the idea, they may not succeed in conveying the complete thought. The presence of well-structured sentences in a review may help the author comprehend the content of a review with ease. Well-structured sentences also indicate to authors that the reviewer put in a lot of thought and effort into writing the review. Similarly in the case of another suggested metric - *word complexity*.

Another metric suggested by a participant is *text cohesion*. Reviews sometimes contain a set of sentences, which may appear to be disconnected, i.e., lack a meaningful flow from one sentence to the next. Cohesive text help make reading and understanding reviews easier.

7.5 Usefulness of the overall automated assessment feature

We surveyed participants on the usefulness of the overall automated feedback system. Out of 24 participants 15 agreed that the feedback was useful (Figure 3), and 8 neither agreed nor disagreed.

One of the participants exclaimed concern with the use of plagiarism as a metric to assess reviews. This is likely because the participant did not see the motivation for a reviewer to plagiarize while writing reviews. Students on Expertiza are evaluated (given scores) on the quality of the reviews they write. Hence they do have a motivation to copy either other good quality reviews (available online) or chunks of text from the submission and submit them as a good quality review. Plagiarism could be caught by manual metareviewers, but may be missed by an automated system. Hence we have this additional feature to ensure that reviewers do not try to game the system by copying reviews.

8. THREATS TO VALIDITY

During the evaluation we noticed that a majority of the participants did not have prior experience in using Expertiza, which could have affected their overall performance.

We also learned, from the comments section of the questionnaire, that a few of the participants did not fully understand the meaning of some of the metrics. An understanding of the purpose of the metareview metrics is essential to assessing their importance and the output's usefulness. Hence, a lack of complete understanding of the metrics may pose as a threat to our results.

No textual reviews were provided by 4 of the participants, which means that the system outputs a value of 0 for each of the metareview metrics. Participants may not be able to discern the usefulness of metrics' outputs for which they have received a score of 0. These are some of the threats to the validity of our results.

9. FUTURE DIRECTIONS

In the future we plan on doing the following: (1) improve the display of metareview output to the reviewer, (2) identify the usefulness of other metareview metrics, (3) study the degree of agreement of the automated metareview ratings with human-provided metareview ratings, and (4) study improvement in reviewing skills.

In order to improve the system's metareview output we plan to highlight snippets of the review that need to be updated. Two participants suggested the need for additional information on metrics such as problem detection and solution suggestion. We plan to provide information on specific instances (of the author's work), which the reviewer needs to read and assess to identify problems or provides suggestions. Also, providing feedback to reviewers with samples of good quality reviews may help them learn how to fix their reviews.

We plan on investigating the use of other metrics such as sentence structure, cohesion and word complexity (discussed in Section 7.4) to study a review's quality. At present our graph-based representations capture sentence structure (e.g. subject-verb-object), but we do not study cohesion across sentences in a review. A study of cohesion may involve exploring other areas of natural language processing such as anaphora resolution [24].

We plan on investigating the extent to which the output from the automated metareview system, as a whole, agrees with human-provided values. This will help us determine whether the system would do as good a job of metareviewing i.e., be as good as human metareviewers in assessing reviews.

We would also like to study if reviewers who get feedback from the system show signs of improvement, i.e., if their reviewing skill improves with time. This would indicate that reviewers learn from the system's feedback to provide more specific and more useful reviews to authors. We would also like to investigate the impact a review quality assessment system has on the overall quality of the authors' submissions.

10. CONCLUSION

Assessment of reviews is an important problem in education, as well as science and human resources, and so it is worthy of serious attention. This paper introduces a novel review quality feature, which uses metrics such as review content type, relevance, tone, quantity and plagiarism to assess reviews. This feature is integrated into Expertiza, a collaborative web-based learning application. We surveyed 24 participants on the importance of the metrics and usefulness of the review quality assessment's output. Results indicate that participants found review relevance to be most important in assessing review quality, and system output from metrics such as review content and plagiarism to be most useful in helping them learn about their reviews.

11. REFERENCES

[1] K. Cho and C. D. Schunn, "Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system," *Computer Education*, vol. 48, pp. 409–426, April 2007.

[2] E. F. Gehringer, L. M. Ehresman, and W. P. Conger, S.G., "Reusable learning objects through peer review: The expertiza approach," in *Innovate: Journal of Online Education*, 2007.

[3] E. F. Gehringer, "Expertiza: Managing feedback in collaborative learning." in *Monitoring and Assessment in Online Collaborative Environments: Emergent Computational Technologies for E-Learning Support*, 2010, pp. 75–96.

[4] R. Rada, A. Michailidis, and W. Wang, "Collaborative hypermedia in a classroom setting," *J. Educ. Multimedia Hypermedia*, vol. 3, pp. 21–36, January 1994.

[5] C. KÃijhne, K. BÃűhm, and J. Z. Yue, "Reviewing the reviewers: A study of author perception on peer reviews in computer science." in *CollaborateCom'10*, 2010, pp. 1–8.

[6] P. Wessa and A. De Rycker, "Reviewing peer reviews: a rule-based approach," in *International Conference on E-Learning (ICEL)*, 2010, pp. 408–418.

[7] J. Burstein, D. Marcu, and K. Knight, "Finding the write stuff: Automatic identification of discourse structure in student essays," *IEEE Intelligent Systems*, vol. 18, pp. 32–39, January 2003.

[8] P. W. Foltz, S. Gilliam, and S. A. Kendall, "Supporting content-based feedback in online writing evaluation with LSA," *Interactive Learning Environments*, vol. 8, pp. 111–129, 2000.

[9] S. van Rooyen, N. Black, and F. Godlee, "Development of the review quality instrument (rqi) for assessing peer reviews of manuscripts," *Journal of Clinical Epidemiology*, vol. 52, no. 7, pp. 625 – 629, 1999.

[10] M. M. Nelson and C. D. Schunn, "The nature of feedback: How different types of peer feedback affect writing performance," in *Instructional Science*, vol. 27, 2009, pp. 375–401.

[11] W. Xiong, D. J. Litman, and C. D. Schunn, "Assessing reviewer's performance based on mining problem localization in peer-review data." in *EDM*, 2010, pp. 211–220.

[12] K. Cho, "Machine classification of peer comments in physics," in *Educational Data Mining*, 2008, pp. 192–196.

[13] R. Zhang and T. Tran, "Review recommendation with graphical model and em algorithm," in *Proceedings of the 19th international conference on World wide web*, ser. WWW '10, 2010, pp. 1219–1220.

[14] S. Moghaddam, M. Jamali, and M. Ester, "Review recommendation: personalized prediction of the quality of online reviews," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, ser. CIKM '11, 2011, pp. 2249–2252.

[15] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2002.

[16] N. Dalvi, R. Kumar, B. Pang, and A. Tomkins, "Matching reviews to objects using a language model," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, ser. EMNLP '09, 2009, pp. 609–618.

[17] I. Titov and R. McDonald, "Modeling online reviews with multi-grain topic models," in *Proceedings of the 17th international conference on World Wide Web*, ser. WWW '08, 2008, pp. 111–120.

[18] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, "Detecting product review spammers using rating behaviors," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, ser. CIKM '10, 2010, pp. 939–948.

[19] L. Ramachandran and E. F. Gehringer, "A word-order based graph representation for relevance identification [poster]," *CIKM 2012, 21st ACM Conference on Information and Knowledge Management*, October 2012.

[20] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and comparing opinions on the web," in *Proceedings of the 14th International Conference on World Wide Web*, 2005, pp. 342–351.

[21] M. Kuniavsky, *Smart Things: Ubiquitous Computing User Experience Design: Ubiquitous Computing User Experience Design*. Morgan Kaufmann, 2010.

[22] E. L.-C. Law, V. Roto, M. Hassenzahl, A. P. Vermeeren, and J. Kort, "Understanding, scoping and defining user experience: a survey approach," in *Proceedings of the 27th international conference on Human factors in computing systems*. ACM, 2009, pp. 719–728.

[23] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[24] E. Tognini-Bonelli, "Corpus linguistics at work," *Computational Linguistics*, vol. 28, no. 4, pp. 583–583, 2002.