# Automated Scoring of Students' Annotations When Learning from Multiple Texts

Alexandra List
The Pennsylvania State University
azl261@psu.edu

## ABSTRACT

Learning about complex and controversial issues often demands that students integrate information from across multiple texts, rather than their being able to rely on only a single resource. This constitutes a highly demanding process. One potential way for students to manage the demands of multiple text learning is through annotation. This secondary analysis of prior work examines whether features of students' digital annotations can be used to classify the types of annotations rendered. Three sets of models were run predicting whether students' digital annotations of multiple texts would be classified as (a) paraphrases, (b) elaborations, or (c) categorizations by expert raters. Models had between 79% and 81% prediction accuracy, and F1 scores greater than .75, suggesting the viability of automated methods to classify students' digital annotations. Moreover, indices of feature importance identified certain features (e.g., annotation length) as particularly valuable for some classification models (e.g., classifying annotations as elaborations or not). Thus, this paper represents an important initial step in developing automated scoring methods of students' digital annotations of multiple texts.

## Keywords

digital annotations, learning from multiple texts, classification.

## 1. INTRODUCTION

Learning about complex and controversial issues, from climate change to immigration, requires that students consult multiple texts, with a single resource often unable to provide students with all of the information they need to understand issues completely and multidimensionally [1]. Yet, learning from multiple texts has also been identified as a highly demanding academic process, requiring that students manage a large volume of information from both complementary and conflicting sources and integrate and corroborate information across texts [2].

One potential way for students to manage the informational and conceptual demands associated with multiple text learning is for them to generate (digital) annotations. Digital annotations refer to students' notes, composed in association with or in response to texts' content, that are co-located or associated with specific content from texts. Digital annotations have three main features [3]: (a) they are responsive to or elicited by the content in texts; thus, students' annotations can be understood as an indicator of texts' perceived noteworthiness or importance; (b) they are generative, reflecting learners' summary, analysis, or critique of texts, with the annotations that students produce differing in their degree of content transformation or prior knowledge engagement and distance from the information provided in-text, and (c) they are co-located or associated with specific text's content; thus, annotations are distinct from linear notes, composed in a stand-alone fashion.

Annotations can serve at least three functions in supporting learning from multiple texts. First, by summarizing or otherwise abbreviating texts' content, annotations can reduce the working memory demands associated with the volume of information introduced by multiple texts. Second, through their co-location with texts' content, annotations can provide a mechanism for students to track concepts across texts; for instance, by explicitly embedding inter-textual connections alongside specific text's content. Third, as is also the case when reading single texts, annotations constitute generative learning outcomes, thus allowing students to externalize their higher-order strategy use (e.g., elaboration, metacognition) in relation to texts' content.

The majority of prior research on annotations has either examined students' annotations of only a single text [4, 5] or students' use of social annotations, modeled by or created in collaboration with others [6, 7], with limited prior work examining students' digital annotations of multiple texts. Prior work has identified a number of categories that may be used to characterize students' annotations when learning from (single) texts. For instance, Yeh et al. (2016) examined students' annotations of weekly articles, read as part of an English as a Foreign Language course, finding these to fall into four main categories. That is, students' annotations focused on (a) *predicting* texts' content, (b) *summarizing*, (c) *clarifying* unfamiliar vocabulary, and (d) *questioning* important information to generate comprehension questions, with students most commonly annotating unfamiliar words [8]. Adams and Wilson (2022) analyzed over 400 social annotations rendered by graduate students, across three class readings, completed throughout the semester. They found these to fall into three main categories, with annotations focused on comprehension, critical literacy, and community [9]. *Comprehension-supporting annotations* including students summarizing, inferencing, monitoring comprehension, and connecting texts to their prior knowledge, to praxis, and to other texts. *Critical literacy-focused annotations* included students pushing back on, or constructing counter-narratives of, texts, engaging in reflexivity or introspection, and focusing on the socio-political context of information. Finally, *community-focused annotations* included students questioning, restating, or adding to others' ideas. Interestingly, the most common type of annotation to emerge was students' formation of text-to-text connections, even though students were only annotating individual texts. This suggests the promise of using digital annotations as a means of fostering students' learning from multiple texts.

A. List. Automated scoring of students' annotations when learning from multiple texts. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 692–697, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12729920

Yet, to my knowledge, using digital annotations to support multiple text learning has only been examined in a limited number of studies. In one unique investigation, List and Lin (2023) analyzed students' digital annotations of multiple texts, when learners were assigned to one of four different task conditions. In particular, students were instructed to either annotate information that was (a) relevant or important in text (i.e., relevance processing condition), (b) related to other texts (i.e., intertextual processing condition), (c) necessary to judge texts' trustworthiness (i.e., evaluation condition), or that was (d) confusing or difficult for students to understand (i.e., metacognitive monitoring condition). Across these four task conditions, five main types of annotation categories emerged [10]. These included students producing annotations that (a) *paraphrased* texts' content, (b) *elaborated* texts' content, (c) *categorized* texts' content (e.g., identifying information as a definition or a statistic), (d) *related* content across texts, or (e) *evaluated* source or content in texts. List and Lin (2023) found that the total number of annotations that students produced predicted both multiple text comprehension and integration performance; although the specific number of annotations, of different types, that students generated was largely not associated with outcomes.

Nevertheless, given prior work emphasizing the importance of examining the types of annotations that students produce [11, 12] and the still limited work investigating digital annotations of multiple texts, this study examines whether the categories of multiple text digital annotations that students produced in List and Lin (2023) are able to be automatically classified. In particular, List and Lin (2023) hand-coded the annotations that students rendered; the present study is a secondary data analyses examining whether researcher-generated annotation categories can be automatically predicted. Being able to automatically predict annotation categories may be a means of further parsing the types of annotations that students render when learning from multiple texts and a precursor to supporting or scaffolding students' production of multiple text digital annotations. Thus, this study has one main research question: *To what extent can researcher-derived categories of scoring students' digital annotations of multiple texts be predicted using automated methods?*

# 2. METHODS

Full methodological information is provided in List and Lin (2023).

## 2.1 Participants

Participants were 278 undergraduate students enrolled in a large university in the United States. The sample was majority (72.94%, n=200) female and majority (74.10%, n=206) White.

## 2.2 Multiple Text Task

Following consent, participants were asked to complete a multiple text task under one of four experimental conditions. The multiple text task had three primary parts. First, participants received task instructions in accordance with their experimental condition (e.g., asking them to annotate important information in text, in the relevance processing condition). Second participants were presented with four texts on the topic of mass incarceration in the United States. Texts were purposefully designed to introduce distinct, complementary, and conflicting content, requiring the formation of inter-textual connections, potentially reflected in the annotations that students produced. Additionally, texts included topic-relevant keywords (e.g., misdemeanor, community corrections), explicitly defined across texts. During reading, students were provided with a highlighting and annotation tool that they could use to annotate each text. Given the co-located or referential nature of annotations, students had to highlight specific portions of text before these were able to be annotated. Finally, after reading all four texts, presented in counterbalanced order, students were asked to write a research report on the topic of mass incarceration as well as to complete objective measures of comprehension and integration.

## 2.3 Annotations

Students' annotations were the primary data source for this study. Students produced a total of 774 annotations across the four texts read. Students' annotations were analyzed across texts and experimental conditions. Thus, here, one specific annotation, rendered in response to one highlighted portion of text, constitutes the unit of analysis. Table 1 includes sample annotations in accordance with categories determined by List and Lin (2023).

**Table 1. Annotation Categories from List & Lin (2023)**

| Category | Sample Annotations | N |
|---|---|---|
| Paraphrase | "Large amount of people in jail" | 165 |
| Elaboration | "Cash bail made it easier for those with money to get out of jail." | 191 |
| Categorization | "Probation definition"<br>"Statistics on arrests" | 228 |
| Intertextual Connection | "Corresponds to the last 3 articles; the amount of people in prisons (A LOT)" | 39 |
| Evaluation | "Dr. Mark Miller studies public policy, which makes the text trustworthy to me." | 98 |

## 2.4 Analytic Plan

Analyses proceeded in four phases. First, annotations, and associated highlighted information from texts, were subjected to standard preprocessing measures. This included (1) removing punctuation and stop words, (2) tokenization, and (3) stemming using the Porter Stemmer algorithm. Second, annotations were explored descriptively (see Figure 1). Third, a number of features (e.g., length, cosign distance) were engineered based on annotation content and the text highlighted in association with each annotation. Finally, the features engineered were used to predict the three most commonly occurring annotation categories (i.e., paraphrases, elaborations, and categorizations).

# 3. RESULTS

## 3.1 Features Engineering

A number of features were created to capture automatically detectable features of students' annotations and the text highlighted in association with each annotation. This included four sets of measures. First, three text-related measures were created. These included (a) the order in which students viewed a text (i.e., given that texts were counterbalanced), (b) the number of keywords within a text that students highlighted, and (c) the number of keywords, from texts, reflected in students' annotations. Second, three length-related metrics were created to capture elaboration. These were (a) the length of students' annotations, (b) the length of the corresponding text that students highlighted, and (c) the ratio of the length of students' annotations to the length of highlighted text. Third, cosine similarity, based on term frequency-inverse document frequency (tf-idf) vectorization, was used as a measure of the semantic overlap between students' annotations and the information highlighted in text. Cosine similarity scores, ranging from 0 to 1, capture the similarity between two documents, in this case students' annotations and highlighted text, with lower cosine similarity values potentially reflecting more-so inferential annotations on the part of learners. Fourth, three parts-of-speech related metrics were created to capture lexical richness. These reflected the percentage of words, within students' annotations, that were (a) nouns, (b) verbs, and (c) adjectives. The full list of features created and used in predictive modeling is included in Table 2. Figure 1 includes a heatmap of the correlations among predictors.

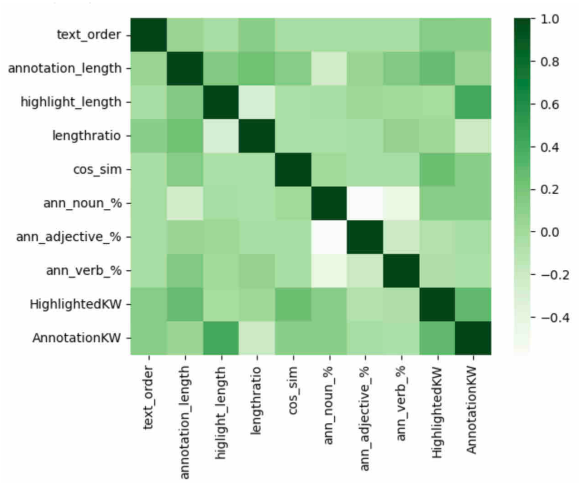

**Figure 1. Heatmap of the Correlations Among Predictors**

**Table 2. Features Used to Predict Annotation Category**

| Feature | M (SD) |
|---|---|
| Order Text Accessed | 3.08 (SD=2.08) |
| Number of KW in Highlighted Text | 0.52 (SD=0.66) |
| Number of KW in Annotations | 1.00 (SD=0.95) |
| Length of Highlighted Text | 99.98 (SD=72.40) |
| Annotation Length | 46.07 (SD=40.21) |
| Ratio of Annotation Length to Highlighted Text Length | 1.42 (SD=4.76) |
| Cosine Similarity between Annotations and Highlighted Text | 0.10 (*SD*=0.16) |
| Percent Nouns in Annotation | 0.61 (*SD*=0.28) |
| Percent Adjectives in Annotation | 0.17 (*SD*=0.21) |
| Percent Verbs in Annotation | 0.11 (*SD*=0.16) |

## 3.2 Classifying Annotations

Three binary outcomes were predicted in this study, corresponding to the three most common annotation categories identified in List and Lin (2023). That is, models were run predicting whether or not students' annotations were classified as (a) paraphrases, (b) elaborations, or (c) categorizations. Four popular classification algorithms were used in predicting each outcome. These were (a) logistic regression (LR), (b) K-Nearest Neighbor (KNN), (c) Random Forest Classifier (RFC), and (d) Gradient Boosting (GB). Grid-search, with five-fold cross-validation, was used for hyperparameter tuning. For each outcome, the KNN classifier was assessed with K = 2, 3, 5, and 10. The RFC was estimated with the number of trees set to 10, 100, 500, and 1000 and maximum tree depth set to None, 3, or 10. The GB Classifier was estimated with boosting stages set to 100, 500, and 1000, and with the maximum depth of regression estimators set to 3, 5, or 10. Data were split into training (70%) and test (30%) sets in each case. Model fit information is summarized in association with each classification category and, where appropriate, weighted averages are computed across classification categories (e.g., paraphrases or not).

### *3.2.1 Predicting Paraphrased Annotations*

Table 3 includes model performance metrics for predicting whether or not students' annotations were classified as paraphrases.

**Table 3. Model Performance Metrics for Classifying Annotations as Paraphrases or Not Paraphrases**

| Algorithm | Accuracy | F1 Score | Precision | Recall |
|---|---|---|---|---|
| LR | 0.768 | 0.679 | 0.674 | 0.768 |
| KNN (K=10) | 0.755 | 0.665 | 0.594 | 0.755 |
| RFC | 0.781 | 0.741 | 0.745 | 0.781 |
| GB | 0.790 | 0.747 | 0.761 | 0.790 |

*Note:* Based on hypermeter tuning the random forest classifier was run with n_estimators = 500 and max_depth = None; the gradient boosting classifier was run with n_estimators = 100 and max_depth = 3.

Comparing algorithms, the gradient boosting algorithm most effectively predicted whether students' annotations were paraphrases or not, with 78.97% accuracy. See Figure 2.

Feature importance values are displayed in Table 4. Cosine similarity and the length of highlighted text and annotation length were the most important features in classifying students' annotations as paraphrases.

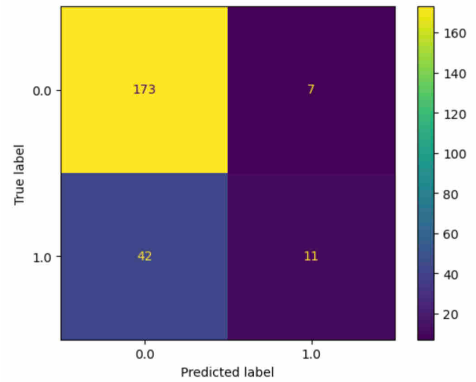

**Figure 2. Confusion Matrix for Annotations Classified as Paraphrases or Not**

**Table 4. Features Importance in Classifying Annotations as Paraphrases**

| Predictor | Feature Importance |
|---|---|
| Cosine Similarity | 0.366 |
| Length of Highlighted Text | 0.158 |
| Annotation Length | 0.122 |
| Length Ratio | 0.108 |
| Percent Nouns in Annotation | 0.062 |
| Number of KW in Highlighted Text | 0.060 |
| Order Text Accessed | 0.049 |
| Number of KW in Annotations | 0.031 |
| Percent Verbs in Annotation | 0.024 |
| Percent Adjectives in Annotation | 0.021 |

### *3.2.2 Predicting Elaborative Annotations*

Table 5 includes model performance indices for classifying students' annotations as elaborations.

**Table 5. Model Performance Metrics for Classifying Annotations as Elaborations or Not Elaborations**

| Algorithm | Accuracy | F1 Score | Precision | Recall |
|---|---|---|---|---|
| LR | 0.777 | 0.750 | 0.748 | 0.777 |
| KNN (K=10) | 0.755 | 0.690 | 0.689 | 0.755 |
| RFC | 0.807 | 0.770 | 0.798 | 0.807 |
| GB | 0.807 | 0.787 | 0.790 | 0.807 |

*Note:* Based on hypermeter tuning the random forest classifier was run with n_estimators = 100 and max_depth = 10; the gradient boosting classifier was run with n_estimators = 100 and max_depth =5.

Comparing performance metrics, both the random forest classifier and the gradient boosting algorithm were effective in predicting whether students' annotations were elaborations or not, with 80.69% accuracy. See Figure 3.

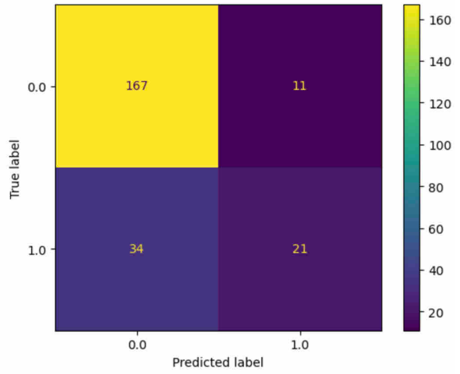

**Figure 3. Confusion Matrix for Annotations Classified as Elaborations or Not**

Features importance values are displayed in Table 6. Annotation length and the ratio of annotation length to the length of highlighted text, both measures of elaboration, were the features most important in classifying students' annotations as elaborations.

**Table 6. Features Importance in Classifying Annotations as Elaborations**

| Predictor | Feature Importance (RF) | Feature Importance (GB) |
|---|---|---|
| Annotation Length | 0.186 | 0.209 |
| Length Ratio | 0.156 | 0.149 |
| Length of Highlighted Text | 0.138 | 0.128 |
| Percent Nouns in Annotation | 0.104 | 0.084 |
| Percent Adjectives in Annotation | 0.076 | 0.080 |
| Cosine Similarity | 0.089 | 0.079 |
| Percent Verbs in Annotation | 0.071 | 0.079 |
| Number of KW in Annotations | 0.052 | 0.074 |
| Number of KW in Highlighted Text | 0.050 | 0.063 |
| Order Text Accessed | 0.076 | 0.054 |

### *3.2.3 Predicting Categorization Annotations*

Table 7 includes model performance indices for classifying students' annotations as categorizations.

**Table 7. Model Performance Metrics for Classifying Annotations as Categorizations or Not Categorizations**

| Algorithm | Accuracy | F1 Score | Precision | Recall |
|---|---|---|---|---|
| LR | 0.785 | 0.769 | 0.774 | 0.785 |
| KNN (K=2) | 0.721 | 0.678 | 0.686 | 0.721 |
| RFC | 0.803 | 0.791 | 0.794 | 0.803 |
| GB | 0.790 | 0.785 | 0.783 | 0.790 |

*Note:* Based on hypermeter tuning the random forest classifier was run with n_estimators = 10 and max_depth = 10; the gradient boosting classifier was run with n_estimators = 100 and max_depth =3.

Comparing performance metrics, the random forest classifier was most effective in predicting whether students' annotations were categorizations or not, with 80.26% accuracy. See Figure 4.

Features importance values are displayed in Table 8. Annotation length, the length of the highlighted text, and the ratio of these two values, as well as the percentage of nouns included in students' annotations, were all important features in classifying annotations as categorizations or not.

**Table 8. Features Importance in Classifying Annotations as Categorizations**

| Predictor | Feature Importance |
|---|---|
| Annotation Length | 0.189 |
| Length of Highlighted Text | 0.164 |
| Length Ratio | 0.160 |
| Percent Nouns in Annotation | 0.139 |
| Number of KW in Annotations | 0.077 |
| Order Text Accessed (1 – 4) | 0.068 |
| Percent Adjectives in Annotation | 0.060 |
| Cosine Similarity | 0.051 |
| Number of KW in Highlighted Text | 0.049 |
| Percent Verbs in Annotation | 0.043 |

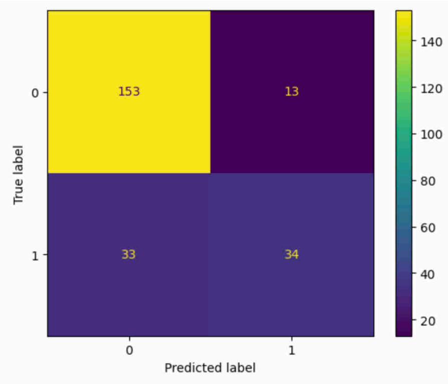

**Figure 4. Confusion Matrix for Annotations Classified as Categorizations or Not**

## 4. DISCUSSION

Digital annotations hold promise for supporting students' learning from multiple texts, a ubiquitous and highly demanding academic outcome [1, 2]. Yet, students' digital annotations when learning from multiple texts have received comparatively little attention in prior research, despite the proliferation of both digital texts and annotation tools. The aim of this study was to examine whether various metrics of students' annotations, able to be generated automatically, could be used to predict how such annotations were classified by expert raters. Three main annotation categories were predicted: (a) paraphrases, (b) elaborations, and (c) classifications. For all three annotation categories, classification models performed quite well, demonstrating between 79 – 81% prediction accuracy, and had F1 scores greater than 0.75, suggesting acceptable to good model performance. This points to the viability of using automated scoring methods for categorizing students' annotations.

Secondarily, measures of feature importance determined that somewhat varied indices were differentially important in classifying annotations into different categories. For instance, cosine similarity, or the overlap between annotations and highlighted text, emerged as the most important feature in predicting whether or not annotations were paraphrases – a logical finding. Likewise, length, or elaboration-related indices, emerged as the most important feature in classifying annotations as elaborations. These results are promising both in suggesting the differentiable nature of annotation categories, identified by expert raters, and in supporting future work seeking to classify students' digital annotations when learning from multiple texts. Thus, this study contributes to the still limited work examining the potential of students' multiple text annotations and identifies promising features to examine in automatically scoring such annotations in future work.

This study has a number of limitations. First, features of the context (e.g., task instructions, texts' content) likely impacted students' annotations but were not explicitly modeled. Thus, examining the types of annotations produced under different task conditions constitutes a direction for future work. Second, students both produced multiple annotations within the dataset and many students participating in the study did not produce any annotations. This requires examining the role of individual differences in students' formation of digital annotations. Likewise, the association between annotations and multiple text task performance requires investigation in future work. Finally, the majority of students participating were White and female. Thus, replicating the study with a larger and more diverse sample is an essential next step.

## 5. REFERENCES

[1] Britt, M. A., Rouet, J. F., & Braasch, J. L. (2012). Documents as entities: Extending the situation model theory of comprehension. In M. Britt, J. F. Rouet, & J.L.G. Braasch (Eds.), *Reading-from words to multiple texts* (pp. 160-179). New York: Routledge.

[2] List, A., & Alexander, P. A. (2019). Toward an integrated framework of multiple text use. *Educational Psychologist*, *54*(1), 20-39. https://doi.org/10.1080/00461520.2017.1328309

[3] Agosti, M., & Ferro, N. (2006). Annotations on digital contents. *International Journal Digital Libraries,* 6(2), 124-138.

[4] Goodwin, A. P., Cho, S. J., Reynolds, D., Brady, K., & Salas, J. (2020). Digital versus paper reading processes and links to comprehension for middle school students. *American Educational Research Journal, 57*(4), 1837-1867. https://doi.org/10.3102/000283121989030

[5] Marshall, C. C. (1997). Annotation: from paper books to the digital library. In *Proceedings of the Second ACM International Conference on Digital libraries* (pp. 131-140). https://dl.acm.org/doi/pdf/10.1145/263690.263806

[6] Sun, C., Hwang, G. J., Yin, Z., Wang, Z., & Wang, Z. (2023). Trends and issues of social annotation in education: A systematic review from 2000 to 2020. *Journal of Computer Assisted Learning*, *39*(2), 329-350. https://doi.org/10.1111/jcal.12764

[7] Wolfe, J. (2008). Annotations and the collaborative digital library: Effects of an aligned annotation interface on student argumentation and reading strategies. *International Journal of Computer-Supported Collaborative Learning*, *3*, 141-164. https://doi.org/10.1007/s11412-008-9040-x

[8] Yeh, H. C., Hung, H. T., & Chiang, Y. H. (2017). The use of online annotations in reading instruction and its impact on

students' reading progress and processes. *ReCALL, 29*(1), 22-38. https://doi.org/10.1017/S0958344016000021

[9] Adams, B., & Wilson, N. S. (2022). Investigating student's during-reading practices through social annotation. *Literacy Research and Instruction*, *61*(4), 339-360. https://doi.org/10.1080/19388071.2021.2008560

[10] List, A., & Lin, C. J. (2023). Content and quantity of highlights and annotations predict learning from multiple digital texts. *Computers & Education*, 199, 104791. https://doi.org/10.1016/j.compedu.2023.104791

[11] Bateman, S., Brooks, C., McCalla, G., & Brusilovsky, P. (2007). Applying collaborative tagging to e-learning. Banff, Canada: *Proceedings of WWW'07.* Retrieved from: https://www2007.cpsc.ucalgary.ca/workshops/paper_56.pdf

[12] Bateman, S., Farzan, R., Brusilovsky, P., & McCalla, G. (2006). OATS: The open annotation and tagging system. Montreal, Canada: *Proceedings of I2LOR.*