Canadian Perspectives on Academic Integrity, 2025 *8*(4), 10-16
http://doi.org/10.55016/ojs/cpai.v8i4/81528
Practitioner Article

Applying Text Analysis for Detecting Academic Misconduct on a Statistics Exam

Douglas Whitaker*

Abstract
This practitioner paper presents an example of using text similarity analysis (specifically using Levenshtein similarity) as one component of an investigation into incidents of academic dishonesty in an online assessment at a Canadian university. The paper begins with an overview of the Levenshtein similarity method followed by a description of the academic offences it was used to provide evidence for. The paper concludes with reflections on the challenges and opportunities the use of text similarity analysis affords.

Keywords
academic dishonesty, Levenshtein similarity, online assessment, text similarity analysis, statistics

***Corresponding author**: Douglas Whitaker, Mount Saint Vincent University, douglas.whitaker@msvu.ca

Introduction

In response to the then-nascent COVID-19 pandemic, Mount Saint Vincent University declared a significant disruption of operations in March 2020. Like many universities, Mount Saint Vincent University courses abruptly shifted to online learning, and many policies were suspended, including requirements for proctored in-person final exams. Text similarity analysis was used to uncover incidents of academic dishonesty on the online, unproctored final exam in an introductory course during the Winter 2023 semester (the last semester of COVID-19-impacted exams). The Levenshtein distance text similarity analysis method was used to provide supplementary evidence for an instance of suspected academic dishonesty as well as to identify other instances of academic dishonesty that would not have been identified otherwise.

Overview of Levenshtein Distance

Text similarity analysis is a broad field of methods used for quantifying the similarity of two (or more) texts. Gomaa & Fahmy (2013) describe two primary classifications of similar text: semantic similarity (i.e., the texts have similar or related meanings) and lexical similarity (i.e., the texts have similar sequences of characters). Methods for semantic similarity analysis require knowledge-based or corpus-based algorithms (Gomaa & Fahmy, 2013) and were not used in this investigation. Methods for quantifying lexical similarity focus on the similarity of strings of text and can be applied either at the term level (e.g., word) or at the character level (Gomaa & Fahmy, 2013). This study used a character-level analysis method: Levenshtein distance.

The Levenshtein distance quantifies the similarity between two strings as the number of edits needed to obtain one string from another (e.g., Zhang et al., 2017). The Levenshtein distance was first proposed in 1965 (Zhang et al., 2017) and has since been used in many applications, including plagiarism detection (e.g., Su et al., 2008). To illustrate the Levenshtein distance, consider the example strings (simulated student responses) shown in Table 1.

The Levenshtein distance between String 1 and String 2, $d(S_1,S_2)$, is determined by the number of edits needed to convert one string to the other. In this case, String 2 can be obtained from String 1 by replacing the capital *B* with a lowercase *b*, replacing the - in p-value with a space, deleting the comma, and deleting the period. Thus, $d(S_1,S_2)=4$. The value of the Levenshtein distance for any string with itself is $d(S_i,S_i)=0$. The values of $d(S_i,S_i)$ for all pairs of the example strings are shown in Table 2; for simplicity of presentation, only the lower triangle of values is shown because the values would be identical in the upper triangle.

This illustrates how the Levenshtein distance can provide meaningful information in the context of plagiarism detection. The maximum value of the Levenshtein distance is related to the length of the longest string compared. Consider comparing String 1 with String 3 and comparing String 5 with String 6. For Strings 1 and 3, the Levenshtein distance is $d(S_1,S_3)=27$. However, for Strings 5 and 6 the Levenshtein distance is considerably smaller: $d(S_5,S_6)=10$. When reading these strings, Strings 1 and 3 clearly have more in common with each other than Strings 5 and 6, but the overall lengths of Strings 5 and 6 limit the maximum of the Levenshtein distance. To address this, the Levenshtein distance is frequently converted to a similarity value (e.g., Sariyar & Borg, 2022):

$$sim(A,B)=1-\frac{d(A,B)}{\max(\text{length}(A),\text{length}(B))}$$

Table 1. Simulated student responses to be used as examples when demonstrating Levenshtein distance and similarity calculations

String ID	String Text
1	Because the assumptions and conditions are met and the p-value is so small (0.0001), we have overwhelming evidence against the null hypothesis.
2	because the assumptions and conditions are met and the p value is so small (0.0001), we have overwhelming evidence against the null hypothesis
3	because the assumptions and conditions are met and the pval = 0.0002 is so small, we have very strong evidence against the null hypothesis.
4	Because the assumptions are met and p-val = 0.0003 < 0.05, we reject the null hypothesis in favour of the alternative.
5	Reject H0
6	I don't know.

Table 2. Levenshtein distances calculated for all pairs of example strings; the upper triangle of values is omitted because it is symmetric to the lower triangle.

	String					
String	1	2	3	4	5	6
1	0					
2	4	0				
3	27	27	0			
4	81	83	77	0		
5	137	135	133	112	0	
6	134	132	130	111	10	0

Table 3. Levenshtein similarity values calculated for all pairs of example strings; the upper triangle of values is omitted because it is symmetric to the lower triangle.

	String					
String	1	2	3	4	5	6
1	1					
2	0.972	1				
3	0.811	0.809	1			
4	0.434	0.411	0.446	1		
5	0.042	0.043	0.043	0.051	1	
6	0.063	0.064	0.065	0.059	0.167	1

This similarity value can also be interpreted as either a proportion or a percentage of similarity, and the value of the Levenshtein similarity for any string with itself is $sim(S_i, S_i) = 1$. The Levenshtein similarity values for the six strings (as proportions and rounded to three decimal places) are given in Table 3.

Observe that when converted to a similarity measure, Strings 1 and 3 are shown to be more similar to each other (81.1%) than Strings 5 and 6 are similar to each other (16.7%) because string length is accounted for.

Application of Levenshtein Similarity

Exam Context

These incidents took place during the unproctored, online final exam in an Introductory Statistics II course. The format of an unproctored, online final exam was initially due to the COVID-19 pandemic and had not reverted to proctored exams yet due to a variety of considerations. This was the last semester of unproctored final exams.

The final exam was administered online through Pearson MyLab. Students had 180 minutes to complete the exam during an 84-hour window. Students were allowed to use the "Save for Later" option, which lets students take breaks rather than completing the exam in a single sitting; with this option, any item that is presented to the student before they use Save for Later cannot be reviewed later by the student. This exam displayed items to students across 16 pages with each page containing related items (a testlet); some pages contained only one item. The exam used multiple choice, short answer (type a number), and constructed response (type a few sentences) items. Instructors manually scored the five constructed response items. The instructional team developed nearly all items on the exam, but a few short items were chosen from Pearson's item bank. Students were not allowed to see their score nor review any items until after the exam period had concluded. This format was similar to the midterm exams from that semester, so students were familiar with the Pearson MyLab system for completing exams.

Application: Supplementary Evidence for Suspected Academic Dishonesty

After the exam window closed, an instructor contacted me because they noticed an uncommon term appearing in responses from 4 of their 32 students. Specifically, the instructor identified four students who used the term "showcase" as in "the graph showcase a normal distribution". This seemed quite

unusual to the instructor and, based on this observation, the instructor closely examined all the constructed response answers from these students. The instructor determined an academic offence was likely to have occurred because the students had very similar answers for several of the constructed response items.

After examining the students' answers to the constructed response items, I agreed that their answers seemed unusually similar for three of the five items. While the initial suspicion of these students was due to the presence of an unusual word, I was unsure of the appropriate weight to place on the unusualness of the term. In that academic year, about 21% of the university's students were international students, and I thought that a group of similar responses stemming from a single unusual word might be attributable to differences in how English was learned, a previous mathematics instructor that the students all had in common, or other cultural reasons. To address this, I chose to focus on the overall similarity of the responses rather than a specific word choice. I sought a way of quantifying the similarity of the responses so I could 1) determine how similar these responses were to each other and 2) determine if the similarity exhibited among these responses was unusual relative to the responses of other students.

I believed text similarity analysis would be well-suited to this situation and investigated possible approaches before deciding to use the Levenshtein similarity. I manually copied all student responses for the three constructed response items from the Pearson MyLab system into a spreadsheet to create a corpus for analysis. A unique code was used for each student so names would not be visible to me during the analysis. I then used the statistical software R (R Core Team, 2023) with the RecordLinkage package (Sariyar & Borg, 2022), which provided a convenient implementation of Levenshtein distance for similarity analysis. I then calculated the Levenshtein similarity between all $\binom{32}{2} - 32 = 464$ pairs of different student responses for each of the three constructed responses items for which there was suspected academic dishonesty. Table 4 provides the resulting Levenshtein similarities. (For the purposes of display, values have been simulated so they approximately match the distribution of the real values.) To help visually identify unusual values, I highlighted values that exceeded twice the mean value; these are shown in Table 4. Minimal example R code demonstrating the use of the RecordLinkage package for reproducing Tables 2 and 3 is included as a supplementary file; this code could be easily modified for use with a spreadsheet of real student responses.

From the similarity values in Table 4, a cluster of five students can be observed: Students 16, 22, 27, 30, and 32 had unusually high Levenshtein similarity values with each other. This process was repeated for the other two constructed response items, and a similar pattern of unusual similarity was observed. Recall that the instructor had initially only suspected four students of academic dishonesty because of a single unusual word. However, using quantitative text analysis revealed that a fifth student's answers were unusually similar to the four suspected students. When shown this information during meetings with the students, each originally denied the accusation, but the first of the five students to complete the test ultimately admitted to sharing photos of the exam with the student's answers. Academic dishonesty charges were pursued for each.

During the analysis, three pairs of students were identified as having unusually similar answers (exceeding the threshold of twice the mean) on exactly one of the other two constructed response questions. After thoroughly comparing the exams of these three pairs of students, these matches were determined to be false positives because the other constructed response answers and overall answers to other items were sufficiently different to support the high Levenshtein similarity value being coincidental. These students were not contacted during the investigation.

Application: Identifying Potential Academic Misconduct

A single instructor raised the initial suspicion about sharing answers, and the above analysis focused on a single section of the course. Because the kind of academic dishonesty detected above could involve students coordinating across sections, I created a new corpus consisting of all responses from all students across all sections to one of the constructed response items. Because constructing the corpus involved tedious copying and pasting, I focused the analysis on the constructed response item with the longest average answers in the prior analysis. This analysis included responses from more than 250 students, resulting in more than 30,000 paired comparisons using Levenshtein similarity; the threshold for flagging was increased to 0.70 for this analysis to reduce false positives. The results of this analysis resulted in the following flags:

- the group of 5 students from the previous analysis,
- a group of 10 students across multiple course sections, and
- 2 pairs of students.

Individual exams were holistically investigated before contacting any of the newly identified students. One of the student pairs had quite similar answers to each other and to an example answer their instructor had provided in class. This pair's answers to other constructed response items were similar to each other but also to examples and definitions given in class; their answers to the other items on the test were also similar, but the students had both done well on this exam and prior exams. Because the similarity seemed to have a common origin of the course materials, this was determined to be a false positive and the students were not contacted.

The group of 10 students had quite similar answers to the constructed response items. Additionally, almost all of these students began the exam at one of two times. Pearson MyLab

Table 4. Levenshtein similarity values calculated for all pairs of example strings; the upper triangle of values is omitted because it is symmetric to the lower triangle.

	Student																	
Student	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
15	1																	
16	0.305	1																
17	0.391	0.370	1															
18	0.244	0.229	0.217	1														
19	0.261	0.429	0.429	0.386	1													
20	0.274	0.232	0.308	0.352	0.266	1												
21	0.195	0.301	0.180	0.341	0.332	0.260	1											
22	0.390	**0.695**	0.281	0.221	0.294	0.354	0.455	1										
23	0.245	0.381	0.315	0.297	0.350	0.197	0.446	0.319	1									
24	0.280	0.249	0.378	0.211	0.376	0.394	0.202	0.354	0.174	1								
25	0.276	0.315	0.456	0.272	0.423	0.390	0.449	0.315	0.31	0.312	1							
26	0.365	0.295	0.216	0.350	0.349	0.262	0.318	0.254	0.363	0.253	0.176	1						
27	0.251	**0.970**	0.290	0.411	0.231	0.278	0.303	**0.684**	0.468	0.303	0.446	0.480	1					
28	0.249	0.330	0.266	0.359	0.404	0.460	0.349	0.279	0.364	0.344	0.220	0.299	0.342	1				
29	0.326	0.248	0.365	0.332	0.339	0.193	0.336	0.446	0.408	0.342	0.076	0.448	0.310	0.209	1			
30	0.269	**0.975**	0.303	0.368	0.308	0.395	0.320	**0.747**	0.257	0.345	0.353	0.279	**0.989**	0.355	0.322	1		
31	0.318	0.405	0.230	0.326	0.302	0.178	0.375	0.232	0.352	0.280	0.406	0.342	0.257	0.340	0.297	0.128	1	
32	0.257	**0.839**	0.320	0.314	0.282	0.348	0.294	0.633	0.355	0.373	0.432	0.465	**0.774**	0.268	0.303	**0.850**	0.201	1

provides the total time in seconds a student has viewed an item for; because students can navigate back and forth across pages, this is only a proxy for determining when a student progresses through each test item. However, for these two clusters of students, an item-by-item analysis of the time spent on each item suggested the students might have been working together on the test simultaneously. The overall similarity of the answers combined with the item timing information was presented to the students, who confirmed they had worked together on the final exam; academic dishonesty charges were pursued for each.

The final pair of students had unusually similar answers for each item that were not similar to any course materials; they did not complete the exam at the same time. These students were presented with the evidence of unusually similar answers, and they admitted to sharing notes (this was an open note test). I then discussed the situation with their instructor, who confirmed the students had a close relationship, and, due to work and family commitments, the students often alternated who attended the class and took notes. While this was an unusual situation, the students' explanation was plausible for no academic offence occurring, and the matter was closed. This was not considered a false positive because of the common source of their answers.

Reflections on Using Text Analysis

The Levenshtein similarity text analysis method was used after this exam in two ways: 1) to provide further clarifying evidence for an instance of suspected academic misconduct, and 2) to serve as preliminary evidence to identify students who may have committed academic misconduct. The Levenshtein similarity between two students' responses was never used as the sole criterion for determining academic misconduct; in both false positive and true positive cases, plausible alternative explanations not involving misconduct were considered. The Levenshtein similarity can be a useful tool in some circumstances for practitioners, though opportunities and challenges exist.

First, the motivation for using a quantitative method to determine text similarity was to provide an alternative form of evidence less influenced by individual instructors' perceptions. Initially, a cluster of responses appeared suspicious due to the presence of an unusual word and the overall similarity among them, raising uncomfortable possibilities. If the use of that unusual word is associated with international student status, the observed similarities may reflect false positives stemming from shared linguistic or cultural backgrounds. Conversely, unusually similar responses from domestic students may go undetected if nothing in their phrasing appears prima facie unusual. This concern is particularly pressing given that international students made up more than 20% of those enrolled at Canadian post-secondary institutions in the 2022–2023 academic year (Statistics Canada, 2024). In such a diverse and multicultural environment, relying on instructors' perceptions of student writing, including what phrasing counts as 'unusual,' may lead to disproportionate focus of academic integrity investigations on international students. Of course, the choice of quantitative methods does not make an analysis free from biases (Sablan, 2019), and care must be taken to ensure equitable outcomes for all students.

String-based text analysis methods are likely to be of diagnostic value when applied to responses within certain length bounds. That is, if Levenshtein similarity was applied to responses consisting of a single word or a short phrase, limited useful information would be uncovered because most responses would be either quite similar (especially correct responses) or dissimilar. Likewise, if Levenshtein similarity was applied to longer responses (e.g., essay length or longer), little additional information would likely be provided: any responses flagged as quite similar would probably also be recognized by a reader (although if responses are read by different readers, string-based methods may provide rudimentary plagiarism checks). For the exam in question, the constructed response item analyzed had responses with a median length of about 550 characters (about a paragraph); the other two constructed response items had median response lengths of about 200 and 400 characters. Two additional constructed response items were included on the test, but these could be correctly answered with much shorter responses and did not seem to be as useful for detecting similar answers. The Levenshtein similarity seemed to perform better with the longer responses than the shorter responses based on false positives. More work would be needed to identify characteristics of responses for which the Levenshtein similarity (or other string-based methods) performed well and performed poorly this will likely vary by discipline and course.

While this analysis was performed an online final exam, the same method could be applied to any assignment that students are not allowed to work together on. In fact, the more student-specific the answer ought to be, the more value there is likely to be in using string-based methods. For example, questions that ask students to interpret or describe a concept in their own words should produce more variation among responses than those that ask students to state a definition or make a short list. Likewise, if formulaic answers or answer templates are provided to students in the course materials, then this common source is likely to obscure inappropriate similarity among answers. In this vein, the choice of flagging thresholds during this investigation (originally twice the mean and later 0.70) was arbitrary and seemed to work well in the context of the student responses observed on this test: by any reasonable measure, the flagged similarity values would be outliers. Different applications of the Levenshtein similarity (and indeed other choices of string-based text analysis method) will warrant different cutoff choices. Practitioner experience in applying these methods should guide the cutoff choices used for investigations, and the people investigating potential academic misconduct must be mindful that exceeding any threshold does not by itself indicate that a student has committed academic misconduct. Text analysis can be a useful source of evidence for one component of an academic integrity investigation; the decision to pursue academic offence charges should always be based on the specific circumstances of the incident.

The Levenshtein similarity worked well for identifying similar responses on this exam in part because of the nature of the cheating that occurred: students who took the exam early in the exam window took photographs of their answers and shared them with other students. When students were retyping the answer, minor differences such as punctuation, capitalization, and spelling often differed even though the overall sentences remained similar – and the Levenshtein similarity quantified this. Other types of cheating such as the use of generative AI for answering questions or even substituting words using a text spinner or thesaurus would not be as easily detectable using string-based methods, but other text analysis approaches could be employed as one source of evidence.

While text analysis is a broad field with great potential for supporting academic integrity, currently available text analysis tools require specialized knowledge and are generally not designed for practitioner use in an academic integrity context. Exceptions exist, such as the cheatR package (Ben-Shachar & Simchon, 2022), which is based on N-gram analysis, but no resources seem to exist to support practitioners in choosing the most appropriate text analysis methods for different types of academic integrity investigations nor for employing existing tools for such purposes. If a growing number of practitioners are to use text analysis methods, educational resources must be available to minimize potential misuse of these tools (e.g., such as interpreting any response that exceeds a threshold as definitive proof of academic misconduct).

Received
May 16, 2025

Accepted
May 16, 2025

Published online
November 12, 2025

References

Ben-Shachar, M. S., & Simchon, A. (2022). *cheatR.* (Version 1.2.1-1) [Computer software]. CRAN. https://CRAN.R-project.org/package=cheatR

Gomaa, W. H., & Fahmy, A. A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications*, *68*, 13–18. https://doi.org/10.5120/11638-7118

R Core Team. (2023). *R: A language and environment for statistical computing.* [Software]. R Foundation for Statistical Computing. http://www.R-project.org/

Sablan, J. R. (2019). Can you really measure that? Combining critical race theory and quantitative methods. *American Educational Research Journal*, *56*, 178–203. https://doi.org/10.3102/0002831218798325

Sariyar, M., & Borg, A. (2022). *RecordLinkage.* (R package version 0.4-12.4) [Computer software]. CRAN. https://CRAN.R-project.org/package=RecordLinkage

Statistics Canada. (2024, November 20). *Canadian post-secondary enrolments and graduates, 2022/2023.* The Daily. https://www150.statcan.gc.ca/n1/daily-quotidien/241120/dq241120b-eng.htm

Su, Z., Ahn, B.-R., Eom, K.-Y., Kang, M.-K., Kim, J.-P., & Kim, M.-K. (2008). Plagiarism detection using the levenshtein distance and smith-waterman algorithm. *2008 3rd International Conference on Innovative Computing Information and Control*, 569. https://doi.org/10.1109/ICICIC.2008.422

Zhang, S., Hu, Y., & Bian, G. (2017). Research on string similarity algorithm based on levenshtein distance. *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference*, 2247–2251. https://doi.org/10.1109/IAEAC.2017.8054419

Appendix

```
# Load the package with the levenshteinDist and levenshteinSim functions
# If you don't have this, install it with:
# install.packages("RecordLinkage")
library(RecordLinkage)
# Create the example strings used in the paper
# These should each be on one line
string1 <- "Because the assumptions and conditions are met and the p-value is so
small (0.0001), we have overwhelming evidence against the null hypothesis."
string2 <- "because the assumptions and conditions are met and the p value is so
small (0.0001) we have overwhelming evidence against the null hypothesis"
string3 <- "because the assumptions and conditions are met and the pval = 0.0002 is
so small, we have very strong evidence against the null hypothesis."
string4 <- "Because the assumptions are met and p-val = 0.0003 < 0.05, we reject
the null hypothesis in favour of the alternative."
string5 <- "Reject H0"
string6 <- "I don't know"
# In practice string_vec would be created by reading in a spreadsheet of responses
# using a function like readxl::read_excel()
string_vec <- c(string1, string2, string3, string4, string5, string6)
sapply(string_vec, nchar) # Info purposes: display numchar per string
# Create empty matrices to store the results
example_distances <-
matrix(NA, nrow = length(string_vec),
ncol = length(string_vec))
example_similarities <-
matrix(NA, nrow = length(string_vec),
ncol = length(string_vec))
for (i in 1:length(string_vec)){
for (j in i:length(string_vec)){
example_distances[j,i] <- levenshteinDist(string_vec[i], string_vec[j])
example_similarities[j,i] <- levenshteinSim(string_vec[i], string_vec[j])
}
}
print(example_distances) # Display raw distance values
# Raw distance not generally needed in practice
print(round(example_similarities, 3)) # Display rounded similarity values
# Now we save the file to a .CSV which can
# be imported to Excel or similar for
# easier visual analysis
# (such as using conditional formatting).
write.csv(example_similarities, file = "example_similarities_output.csv")
```