Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Summer 2021, Vol. 40, No. 2, pp. 72–80

A Rubric for the Detection of Students in Crisis

Amy Burkhardt, *University of Colorado at Boulder*, Susan Lottridge, *Cambium Associates*, and Sherri Woolf, *American Institutes for Research (AIR)*

Abstract: *For some students, standardized tests serve as a conduit to disclose sensitive issues of harm or distress that may otherwise go unreported. By detecting this writing, known as* crisis papers*, testing programs have a unique opportunity to assist in mitigating the risk of harm to these students. The use of machine learning to automatically detect such writing is necessary in the context of online tests and automated scoring. To achieve a detection system that is accurate, humans must first consistently label the data that are used to train the model. This paper argues that the existing guidelines are not sufficient for this task and proposes a three-level rubric to guide the collection of the training data. In showcasing the fundamental machine learning procedures for creating an automatic text classification system, the following evidence emerges as support of the operational use of this rubric. First, hand-scorers largely agree with one another in assigning labels to text according to the rubric. Additionally, when this labeled data are used to train a baseline classifier, the model exhibits promising performance. Recommendations are made for improving the hand-scoring training process, with the ultimate goal of quickly and accurately assisting students in crisis.*

Keywords: standardized tests, crisis papers, machine learning, automated scoring, mental health

As the number of reported cases of mental health issues in U.S. children continues to grow (Centers for Disease Control and Prevention, 2019), the suicide rate for teenagers in the United States has increased 25% from 2016 to 2019 (America's Health Rankings, 2019). These recent figures should prompt a deeper understanding of the various facets of young people's lives that might be contributing to this degradation in mental health. A recent investigation, for example, posits several reasons why school may be more stressful now than ever before: Lockdown drills in the wake of recent school shootings, a new form of academic competition enabled by publicly broadcasting performance on social media, forecasts of a grim future if not accepted into a four-year college (which is met by the paradoxical concern of ever-rising tuition costs), and an uptick in hours spent preparing for, and taking tests for which there are consequences for both teachers and students (Brundin, 2019). Though its lasting impact is not yet known, the recent pandemic of COVID-19 will likely be a source of additional stress in students' lives. Students are now responding to the abrupt upending of their social communities within classrooms, without any certainty as to when, if ever, school will return to normal; they must adapt to learning in isolation, and for many, this new learning environment further exacerbates issues of inequality. Even though the solutions to these problems may not be readily apparent, it does seem evident that there are a number of issues the education community can address to alleviate stress and improve the quality of student's lives. This paper highlights one unexpected instrument that can assist in providing mental health support for students: the standardized test.

That is, while tests can invoke stress, they can also serve as a unique opportunity for students to, in effect, cry out for help. Within a test, students may disclose a number of issues, such as anxiety, depression, suicide, other forms of self-harm, and abuse. Though this is not a common practice for students, this form of student writing occurs frequently enough to be recognized by some testing programs as *crisis papers*. One reasonable explanation for this unusual use of tests is that the subject matter of these reports is oftentimes difficult to talk about, and students may find it easier to disclose sensitive information indirectly through writing, rather than verbally and directly to an authority figure. It follows then that, if these self-disclosures are detected, then students who may otherwise be ignored may instead receive the support they need.

Historically, the responsibility of detecting crisis papers has been assumed by hand-scorers as an ancillary part of the scoring process. In the event that a hand-scorer reads a piece of text that contains writing that invokes alarm, the response is flagged, and is routed to the appropriate school personnel to further assess the writing and determine if intervention with the student is necessary.

However, in the digital age, this process must evolve in response to two converging technological developments. First, tests are now commonly delivered on an online platform, and they are often equipped with a digital notepad for drafting answers to test questions. Students may use these virtual pieces of scratch paper to write down candid thoughts and self-reflections, and even though these writings do not receive a score, they arguably should still be reviewed to

© 2020 by the National Council on Measurement in Education

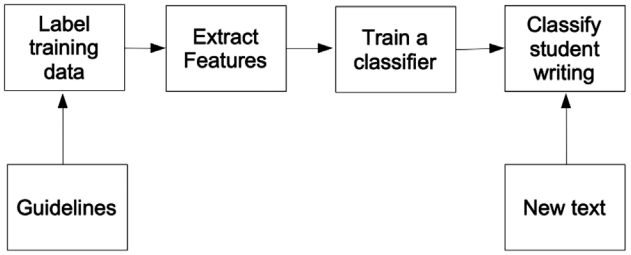

FIGURE 1. Key machine learning components of an automatic text classification system.

ensure that the student is not reporting harm. This, however, places a strain on grading time and resources. Furthermore, as automated scoring becomes more integrated into assessment programs, a manual review of student work is, in some instances, either not part of the process at all, or only consists of a sample of all responses. Therefore, one solution to ensure that testing programs implement a feasible system to detect crisis papers in a timely manner is to transition the task of detecting crisis papers from humans to a machine. Not only does this solution address the technological advances related to testing, but an automated model also offers the unique affordance of consistently detecting crisis papers, without fatigue or distraction.

Even though automated scoring systems and crisis detection systems serve different objectives in evaluating student work, these automated classification systems rely on the same underlying machine learning procedures. Figure 1 presents the key machine learning procedures for creating either one of these automatic classification systems.

For both tasks, the first step is to identify examples of student writing that can then be used to teach a model how to classify text. Using machine learning terminology, this is referred to as labeling the training data. These training data consist of both the outcome variable (the *labels*) and predictor variables (or *features*). While features can be extracted from the text through a number of different procedures, one way is to simply treat each word within the training data as a predictor variable used to predict the outcome (of course, the words must be transformed into numeric representations prior to fitting a model; this process is described in more detail in a subsequent section). By means of a loss function, the classifier learns how to best assign weights to these features in order to predict the outcome value. Then, text that was not used to train the classifier can be passed through the calibrated model to be automatically classified. While decisions are made at each step of the process to improve the model's accuracy, the importance of the first step of the process, of obtaining labeled training data, cannot be overstated. This step is inherently error prone, as humans are fallible and training data will never be labeled with perfect consistency or accuracy. As such, procedures should be implemented to support humans in this task.

For this first step, an automated scoring system relies on data that have been labeled in accordance with a rubric, which serves to delineate the criteria for the different score points and guide the hand-scorers to accurately and consistently label data. In the parallel development of a crisis detection system, there is no industry-established approach for labeling these training data; there is no equivalent of a rubric to guide hand-scorers in labeling data. This is not to suggest that there are not training and monitoring efforts in place to ensure the detection of crisis papers using existing guidelines. The Smarter Balanced Assessment Consortium, for example, instructs hand-scorers to flag text that is suggestive of (but not limited to) the following forms of harm: suicide, criminal activity, alcohol or drug use, extreme depression, violence, rape, sexual or physical abuse, self-harm or intent to harm others, or neglect (Smarter Balanced Assessment Consortium, 2014, p. 10).

However, given the fallible nature of humans, a rubric that more clearly and procedurally guides the hand-scorers through the classification of student writing has potential to improve the quality of the labeled training data. Such data are required for a high-quality detection system, which must attend to two different forms of misclassifications: false positives (i.e., text that is normal but is misclassified as a crisis paper) and false negatives (i.e., text that is actually indicative of harm, but has not been classified as such). On the one hand, it is important to minimize the number of false positives so that the detection system can be both a credible and useful tool. On the other hand, it is critical that the system errs on the side of the health and safety of students.

When classifications are based on guidelines such as those used by the Smarter Balanced Consortium that adopt broad criteria so as to not omit possible instances of harm, it seems likely that false positives will be more prevalent than false negatives. These false positive misclassifications can be broadly categorized into two distinct types of student writing. The first is student writing that is concerning in nature but does not constitute harm. Such text likely exhibits indicators of poor mental health; a student might describe, for example, that they are lonely, or depressed, but without actually stating any threats of self-harm. Even though the writing does not meet the criteria of a harmful situation, the situation has potential to become harmful. In such cases, a hand-scorer might be hesitant to ignore a veiled request for help, and misclassify the student writing, either because of a misinterpretation of the guidelines, or as a deliberate and conscientious objection, in the belief that the guidelines are not in the best interest of the student. Given the current mental health crisis among students, this type of writing should arguably be flagged, a viewpoint that is supported by the best practice guidelines for large-scale assessment (Council of Chief State School Officers and Association of Test Publishers, 2010). Yet, it is also arguable that such text should not be classified as the same severity as students' reports of immediate harm. The second type of student writing that may be likely to be misclassified as a false positive is text that contains any of the following: unusual language, test-related complaints, or test-related requests for intervention. Such text may cause a hand-scorer to give pause as whether or not to flag the text and may ultimately be misclassified due to a misinterpretation of either the student writing or the guidelines (but likely not in a deliberate overruling of the guidelines).

In order to improve the accuracy of the detection of crisis papers, this paper introduces a rubric that is intended to be an industry-standard for labeling training data. This proposed rubric consists of three levels to guide the labeling of three types of student writing. The first level consists of normal writing, in the sense that these pieces of text do not raise concerns about a student's well-being; the second level consists of writing that is concerning in nature and might require intervention with the student; and the third level consists of student writing that indicates harm and should

© 2020 by the National Council on Measurement in Education

immediately trigger the protocol to alert the relevant personnel at the state department or school district to follow-up with the student. These three types of writing that characterize each level of the rubric will subsequently be described using the following shorthand: *Normal (Level 1)*, *Concerning (Level 2)*, and *Alert (Level 3)*. The use of the word "normal" is intended to characterize student writing that is typical of what is written on tests and is not intended as rhetoric to imply or reinforce stigma associated with mental health issues. This three-level design also offers assistance to states and school districts by providing an initial triage of the severity of harm, so that they can prioritize their allocated resources to intervene first with the highest risk students.

This paper also showcases the application of the rubric in the context of stepping through each of the key machine learning components of an automatic text classification system outlined in Figure 1. Through this demonstration, the extent to which the rubric can be consistently applied by hand-scorers is evaluated, as is the extent to which a model's classifications agree with the resolved hand-scorer's designations. The objectives of this paper are organized into two sections: The first section documents the rubric's development, while the second section illustrates the use of this rubric in the context of creating a text classification system.

Developing the Crisis Paper Rubric

The rubric was developed through the following four steps. First, the analytic sample was automatically created by classifying thousands of pieces of student writing using a previously trained model. Second, the contents of the analytic sample were discretely categorized and summarized by applying qualitative codes to the text. Third, in consultation with a hand-scoring director, each of the categories was mapped to one of the three levels of the rubric. Finally, the categories were further aggregated to produce the specific criteria of each level of the rubric.

The Analytic Sample

In order to adequately delineate the criteria for each level of the rubric, the analytic sample needed to include a wide range of student writing. As such, the sample was automatically created by using a model that had been previously trained on data labeled in accordance with the Smarter Balanced guidelines. This model was problematic for operational use because it classified too many responses as crisis papers; however, it was ideal for this context. Not only did the model classify text that was truly indicative of harm as crisis papers, but it also classified text that contained concerning language as well as text that was benign in nature. This model classified student writing of all grade levels, both in ELA and math, from interim and summative tests whose online test administration was supported by the American Institutes of Research[1] in the 2017–2018 academic year. Test administration data came from more than 10 states. Student writing from all text fields within the test were included in this analysis, which included answers to items as well as text located within the virtual notepads.

From over 41,000 responses, the model classified 8,284 responses as *alerts*.[2] However, when hand-scorers subsequently reviewed these classified texts, they identified that 493 pieces of writing were aligned with the Smarter Balanced guidelines, leaving the other 7,791 responses to be considered *non-alerts*. This set of results reinforces the earlier notion that a model trained in accordance with hand-scoring guidelines would result in a large number of false positive classifications. Additionally, during this manual review, hand-scorers also identified 40 responses that were classified as non-alerts to be deemed as alerts and were included in the analytic sample. In all, 8,324 pieces of student writing comprised this sample, with 7,791 classified by hand-scorers as non-alerts and 533 as alerts. All pieces of student writing were 10 words or shorter. Shorter texts afforded the opportunity to review a larger variety of student writing. This decision to only review shorter text invites the concern that an analytic sample that contained longer pieces of writing may have a different impact on the rubric development. This would especially be concerning if longer crisis papers differed in language or content, compared to those of shorter papers. An initial inspection of a larger pool of lengthy papers did not appear to support this notion. A more in-depth exploration of the implications of the decision to only review shorter texts is discussed later in the paper, as text of any length was used in the exercise of applying the rubric.

Qualitative Coding Procedure

The text within the analytic sample was then categorized by means of a qualitative analysis technique of iteratively and systematically reviewing and assigning codes to each piece of student writing in which each code (a word or a phrase) represented a characteristic or phenomenon of interest in the data (Corbin & Strauss, 2015). The coding scheme was composed of three distinct sets of codes: (1) codes capturing reports of imminent, past, or current harm, (2) codes representing negative emotions, and (3) codes describing responses that do not exhibit a clear instance of harm or emotion.

The codes capturing harm were intended to identify all of the various forms of harm that should be included in *Alert (Level 3)* of the rubric. While these were largely a predefined set of codes, additional codes were developed, based on patterns that emerged from the data. The initial codes characterizing reports of imminent, past, or current harm were informed by the Smarter Balanced guidelines: *suicide, criminal activity, alcohol or drug use, extreme depression, violence, rape, sexual abuse, physical abuse, self-harm or intent to harm others*, and *neglect*.

The codes representing negative emotions were included based on the hypothesis that student writing that is indicative of potential harm, or a *Concerning (Level 2)* text, would oftentimes be characterized by the emotion the student conveyed. Take, for example, when students use nonliteral but violent phrases and hostile language to express frustration over a test, or when a student conveys feelings of inadequacy about not measuring up to academic expectations. Students

[1]The Assessment Division of American Institutes for Research was acquired after the date of this study by Cambium Learning Group.

[2]The sample of 41,425 student responses was taken from a larger sample of 850,000 responses that were flagged as alerts by a previous implementation of the model that classified 85 million pieces of student writing. For this study, this sample of 850,000 responses was further restricted to be unique pieces of student writing, consisting of 10 words or less.

also make use of other emotive language, such as sadness or loneliness, that express negative sentiment toward themselves and their lives and, while hinting at a request for intervention, they do not explicitly report imminent harm. The initial codes capturing emotion relied on a list of 19 negative emotions: *anger, anxiety, blame, disappointment, disgust, embarrassment, fear, frustration, grief, guilt, humiliation, hurt, jealousy, loneliness, feeling overwhelmed, regret, sadness, shame*, and *worry* (Brown, 2018).

Finally, because not all writing could be characterized as either harmful or emotionally derived, codes unrelated to either were also included. Examples of these nonemotive codes that emerged from the data included: *physical discomfort*, *language issues*, *reports of not having opportunity to learn the material*, catch-all category of *other* for written work that could not be categorized by any of the other codes.

For the non-alert text, 50% of the 7,791 papers were placed into the catch-all category of *other*. These responses were typically characterized as being considered written on-topic for a prompt, but the subject matter of the prompt contained language that was also sometimes associated with crisis papers (such as on the topic of the *death*). Of the remaining pieces of non-alert text, 70% were characterized by an emotion code. Of the original 19 emotion codes, 11 were used at least once: *anger, anxiety, fear, frustration, grief, hurt, jealously, loneliness, overwhelmed, sadness*, and *shame*. Four additional emotion codes were added based on what was present in the data: *apathetic, bored, confused*, and *emotional fatigue* (i.e., a student is emotionally exhausted and reports that they cannot continue on with the test). Of these, *frustrated* was the most common emotion, followed by *anger, confusion, being overwhelmed, apathetic*, and *feeling shame* (i.e., a student negatively compares the self's actions with the self's standards). Two frequent nonemotion codes characterized when a student simply stated that they did not know the answer, and when they reported complaints of physical discomfort.

Of the 533 alert responses, only 40% of the pieces of text were assigned a harm code, while the other 60% of the sample was composed of the emotion codes of *frustrated*, *overwhelmed*, *anger*, *shame*, *bored*, and *sad*. This finding further supports the claim that, within the framework of the "status quo" flagging approach, hand-scorers will inconsistently classify text that is concerning, but not explicitly harmful.

Mapping Codes to Levels and Finalizing the Rubric

Each of these codes were then mapped to a level of the rubric. The mapping process was performed with the consultation of a hand-scoring director, who had 25 years of experience managing and directing hand-scoring centers and overseeing the assignment of scores to millions of pieces of student writing. The alignment between the codes and the levels was thoroughly reviewed by the hand-scoring director to ensure that these categories, and ultimately the rubric, would not be problematic in operational use. In other words, this step was implemented to avoid two pitfalls that may come with developing a rubric without the institutionalized knowledge of a hand-scoring center. First, it was important to ensure that the rubric would not cause confusion to the hand-scorers, and that the rubric and categories were developed in a manner that was consistent with how rubric information is typically communicated in a hand-scoring center. Second, it was also important to ensure that the criteria for what constituted an *Alert (Level 3)* or *Concerning (Level 2)* paper would not overwhelm the hand-scoring center by flagging responses that were arguably normal in the context of student writing on standardized tests.

These qualitative codes continued to be refined as they underwent the mapping process, in which each qualitative code was linked to one of the three levels. For example, the code *Frustrated* was further refined into two subcodes: *Frustrated: Hyperbolic (Frustrated statements that include violent language)*, which was mapped to *Concerning (Level 2)*, and *Frustrated: Irritated (Characterized by being annoyed or irritated)*, which was mapped to *Normal (Level 1)*. The final coding scheme and each code's location within the rubric levels is presented in Table 1, and the abridged version of this table, which is the final rubric, is presented in Table 2.

Table 1 indicates that all of the harm codes were mapped within the *Alert (Level 1)* of the rubric. The emotion codes were oftentimes mapped to *Concerning (Level 2)* of the rubric, although not always. The emotion code representing *Sad: Extreme* was included in Level 3, in which student disclosed signs of depression or self-hatred that were explicit enough to suggest that self-harm was imminent. Most of the *Concerning (Level 2)* codes are emotive in nature, with the exception of three: (1) a nonthreatening mention of a gun, (2) text that gives pause to the reader that a student is the perpetrator or victim of violence without any clear explicit wording (which includes violent or incoherent language and hints that something may be wrong), and (3) text that contains sexual imagery that may allude to a problematic sexual encounter without directly threatening or reporting abuse. Emotion codes were also mapped to *Normal (Level 1)*, which oftentimes included irritation or anger toward the test. In addition, nonemotive codes, including technological issues with the test, and reports that the student did not know the content, due to an opportunity to learn or otherwise, were also mapped to this first level.

These codes were then further aggregated to create categories within each level. For example, the emotions of sad, lonely, hurt, grief, anxiety, and fear were used to describe a subcategory within *Concerning (Level 2)*: "Signs of depression, self-loathing, or anxiety." The final rubric is presented in Table 2 and was used, in conjunction with examples of writing at each level, to train the hand-scorers how to differentiate between these three types of student writing. In the next section, an overview of this training process is provided.

Applying the Rubric to the Context of a Machine Learning Model

The remainder of this paper applies the crisis paper rubric in the context of building a machine learning model, by first training hand-scorers to label thousands of pieces of student writing, and then using these data to train a classifier to automatically detect crisis papers. Recall from Figure 1 that the machine learning process broadly consists of four machine learning procedures for text classification: (1) label training data, (2) extract features, (3) train a classifier, and (4) classify text that was not used to train the model. While this overview remains at a conceptual level, see Raschka and Mirjalili (2019) and Jurafsky and Martin (2009) for a more comprehensive introduction to machine learning and natural language processing, respectively.

© 2020 by the National Council on Measurement in Education

Table 1. Final Qualitative Coding Scheme for Normal, Concerning, and Alerts

Qualitative Code	Description
Level 3: Alerts[a]	
Sad: Extreme	Expressing extreme signs of depression
Suicide	Describing suicidal ideation or attempt
Violence	Reporting or threatening violence
Rape	Reporting or threating rape
Abuse	Reporting or threating abuse
Drugs	Reporting or using drugs
Help: Specific	Specific and serious request for help (not test-related)
Level 2: Concerning	
Overwhelmed	Request for help that is not specific, nor test-related
Frustrated: Hyperbolic	Frustrated statements that include violent language
Anger: Hate (protected class)	Derogatory terms used to attack protected class (hate speech)
Hurt	Reports dissatisfaction over quality of life
Sad	Express signs of depression, unhappiness, or need to cry
Lonely	Lack of social support, feelings of being isolated, or abandoned
Grief	Express loss of relationship
Anxiety/fear	Express anxiety, fear, or stress
Shame	Negatively compares self's actions with self's standards
Gun**[b]	Nonthreating mention of a gun
Suggestive[b]	Text suspect of student being the perpetrator or victim of violence
Sexual[b]	Sexual imagery without threatening abuse or reporting abuse
Level 1: Normal	
Frustrated: Irritated	Characterized by being annoyed or irritated
Anger: Hate	Hating someone or something without threats of harm or violence
Anger: Disparaging	Disparaging remarks and/or other interpersonal distress
Anger: Jealousy	Jealous remarks indicating interpersonal distress
Bored	Expressing boredom
Apathy	Expressing lack of interest in test
Emotional fatigue	Displays of emotional fatigue
Confused	Expressing confusion regarding test
Help: Test[b]	Test-related request for help
OTL[b]	Reports not having opportunity to learn
IDK[b]	Reports not knowing content
Technical[b]	Technical issues with computer and test
Other	Normal responses that do not fall into any of the above categories

[a] The only emotive code for alerts is identifying cases of extreme depression (Sad:Extreme).
[b] Nonemotive codes.

Table 2. Crisis Paper Rubric

Level	Category	Details
3 - Alert		
	Harm to self or another being	Suicide, self-harm, or extreme depression; threats or reports of violence, rape, abuse, drug use, eating disorders, or neglect; hate speech with threats of violence
	Contains mention of a gun	Must be threatening
	Specific and serious request for help	Not test-related
2 - Concerning		
	Non-specific request for help	Not specific to test or harm
	Sexual imagery	Without threats or reports of abuse
	Violent words or phrases	No explicit reports of being the perpetrator or victim of violence, but text seems suspect; nonthreatening mention of a gun; hate speech without threats
	Signs of depression, self-loathing, or anxiety	Sad, lack of social support, dissatisfaction for life, grief, anxiety, negative attitude toward self. Includes hyperbolic language about wanting to die
1 - Normal		
	Hating another entity	Hate toward someone or thing, and without threats of harm or violence. Does not include hate speech
	Situational testing issues	Complaints of physical discomfort, technical issues with test, not understanding English, lack of engagement with the test, not knowing what to do on the test, frustration with the test
	All other responses	All responses that are not characterized by any of the criteria outlined in this definition

Table 3. Cross Tabulation between Hand-Scorers (Quadratic Weighted Kappa = .87)

	Scorer 2			
Scorer 1	**Alert**	**Concerning**	**Normal**	**Total**
Alert	2,391 (19%)	401 (3%)	83 (1%)	2,875
Concerning	464 (4%)	1,413 (11%)	289 (2%)	2,166
Normal	102 (1%)	317 (2%)	7,457 (58%)	7,876
Total	2,957	2,131	7,829	12,917

Note: Percentages are computed as the proportion of the grand total.

Labeling the Training Data

The training data must contain an abundance of examples of student writing, because students use a variety of different words to express themselves. Therefore, approximately 13,000 unique pieces of student writing across grade levels were selected for the task of locating student writing in the three levels of the rubric. Out of these, 5,000 (38.5%) were confirmed as crisis papers, according to hand-scorers who followed the existing Smarter Balanced guidelines. One hypothesis of this study was that the labels of these texts would likely be redistributed between the levels of *Alert (Level 3)* and *Concerning (Level 2)* when scored according to the newly developed rubric. Eight thousand pieces of student writing that were considered benign were also included. Since this task focused on exercising the different levels of the rubric, the normal responses of interest were those that would be likely boundary cases between *Normal (Level 1)* and *Concerning (Level 2)*.

Six hand-scorers with experience grading open-ended assessments were selected for the study. In order to assess the extent to which hand-scorers with no prior exposure to a crisis detection system could consistently apply the rubric, the only background any of the scorers had with detecting crisis papers would have been as part of the ancillary work of routine grading procedures. Supplementary training resources were created, which included examples of student work at each level of the rubric, and three qualifying sets, consisting of 10 examples of student writing, to ensure that the raters were scoring as intended during the training process. These qualifying sets were administered and graded after the hand-scoring director facilitated a thorough review of the rubric. For each qualifying set, the lowest score was 80% for a hand-scorer, and the modal score was 90% correct. These results were deemed sufficient to qualify each of the raters to proceed in the process. Three examples of student writing were administered daily and graded by the lead hand-scorer to ensure that the hand-scorers continued to be consistent with the rubric. For six days, the hand-scorers labeled papers, and in the end, the exact agreement between two hand-scorers was 87%, adjacent agreement was 11%, and nonadjacent agreement rate was 1.4%. Table 3 presents the cross-tabulation.

Text labeled as *Normal (Level 1)* was the easiest for two hand-scorers to agree on (exact agreement rate of 95%), followed by text located in the *Alert (Level 3)* category, where hand-scorers agreed on 83% of the placements. Text identified as *Concerning (Level 2)* exhibited the most issues in agreement (exact agreement of 65%).[3] These hand-scoring discrepancies are described in more detail in the Discussion section. Discrepancies between raters were reviewed and resolved by the lead hand-scorer. These resolved labels were then used to train the model in the next section.

As expected, after applying the rubric to these pieces of text, the frequency of student writing that was previously labeled as *Normal (Level 1)* stayed relatively the same, but a significant proportion of writing that had been labeled as an *Alert (Level 3)* changed to *Concerning (Level 2)*. The new distribution of labels was 61% *Normal (Level 1)*, 16% *Concerning (Level 2)*, and 23% *Alert (Level 3)*. The barplot in Figure 2 depicts the redistribution of labels.

These newly labeled data were then randomly divided into two samples. Eighty-percent of the data were used to build the model (sample 1), while the other 20% were set aside from the training process and were used as the test set to evaluate how well the model would classify student writing that was not used to train the model (sample 2).

Building and Evaluating a Text Classification Model

To extract features, student writing was transformed into vectorized representations of the text. The most basic version of this method treats all words (i.e., "terms") in the training sample as dummy variables (equal to 1 if word appears in the student writing, 0 otherwise). A vector of ones and zeros is created for each piece of student writing, or "document," in the training sample. An alternative approach is to count the frequency that each word appears in the document, rather than using a binary indicator, which is the approach used in this study. The end result is a matrix that is referred to as a *term-document* matrix.

In this particular term-document matrix, there were over 37,000 terms and approximately 13,000 documents. To reduce the possible noise from the words in a student response that may not be useful in classifying student writing in accordance with the crisis paper rubric, a decomposition of variance technique was applied, called latent semantic analysis (LSA), to reduce these 37,000 words down to 3,500 components.[4] LSA has been used in the context of automated scoring (Foltz et al., 1999) and other intelligent tutoring systems (e.g., Dessus et al., 2000; Kintsch et al., 2000), and is thought to reflect the salient semantic information of a text document.

The features, in the form of an LSA matrix, were then used as input into a classifier, where each component of the LSA matrix contributed to the prediction of whether or not a piece of text should be considered *Normal (Level 1), Concerning (Level 2)*, or *Alert (Level 3)*. While such features can serve as input for a number of different classifiers, this demonstration

[3]These values represent the row percentages, which is the agreement of Scorer 2 conditioned on Scorer 1.

[4]The number of components was selected through a series of small experiments using the training data to determine the optimal number of components.

© 2020 by the National Council on Measurement in Education

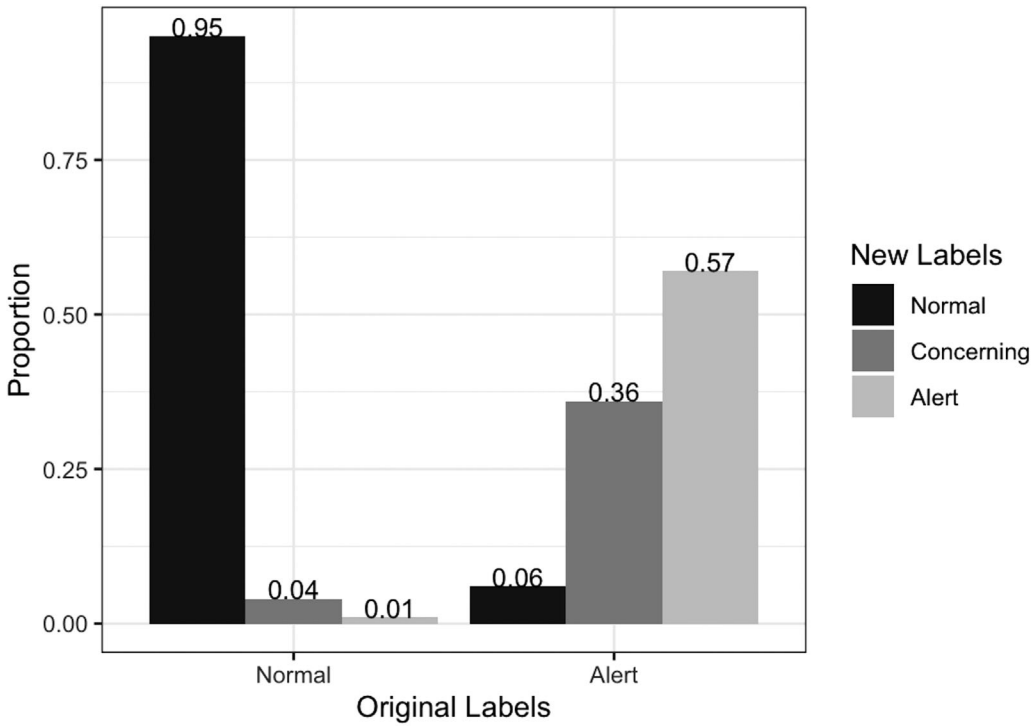

FIGURE 2. The redistribution of the original labels by hand-scorers in accordance with the three-level rubric.

used a feed forward neural network.[5] Briefly, the architecture of the neural network included an input layer consisting of the 3,500 features of the LSA matrix; three hidden layers, each with a different number of hidden units (1,000, 250, and 50); and an output layer that produced a probability distribution across the three crisis paper categories. Each piece of student writing was classified as belonging to the category with the highest probability. The loss function of categorical cross-entropy was applied to minimize the difference between the predicted class probabilities and the resolved label.

Recall that sample 2 was set aside and not used in the training process. Once the classifier had been trained, these data were used to validate the model by assessing how the model performed at classifying previously unseen student data. Table 4 presents these validation results. In this table, the distributions of labels are quite similar for both the classifier and the resolved hand-score, with both systems assigning a similar proportion of *Alert* (23%), *Concerning* (16%), and *Normal* (61%) labels. The exact agreement rates, as well as the adjacent agreement rates, are similar for both hand-scorers and the agreement between the model's classifications and the resolved labels. However, there is a higher non-adjacent rate for the model (3%), compared to the hand-scorers (1%).

Table 4. Validation Results from Hand-Scores and Neural Network (n = 2,584)

	Hand-Scores	Neural Network
Label distribution (%)		
Alert	23	23
Concerning	16	16.5
Normal	61	61
Agreement (%)		
Exact	88	86
Adjacent	11	11
Nonadjacent	1.0	3.0
Quadratic weighted kappa	.89	.84

Note: The hand-scores label distribution is based on the resolved score. The hand-score agreement rate uses the two hand-scorers, and the agreement rate for the neural network relies on the resolved score.

Discussion

This paper introduces a three-level rubric that can be used to train data for the automatic detection of crisis papers. The rubric was developed, in part, by a systematic review of thousands of pieces of student writing, as an attempt to capture the wide range of ways that students may express various levels of distress. The practicality of using this rubric operationally was supported by the following two pieces of evidence: first, in classifying 13,000 pieces of student writing, hand-scorers demonstrated that they were able to consistently apply the criteria of the rubric. This resulted in labeled data that could be used as input for an automatic detection system. Second, validation results from the LSA input into the neural network model demonstrated that the labeled data could be used to effectively train a classifier to automatically detect crisis papers.

Even though the model in this study meets the performance standards for automated scoring (Williamson, Xi, & Breyer, 2012), the possible consequences of misclassifying a piece of text in which a student is asking for help demand stricter standards. As such, identifying ways to improve model accuracy is important. The classification accuracy of a model

[5]For a primer on using neural networks for natural language processing tasks, see Goldberg (2017). For a general overview of a number of different classifiers that can be used for classification tasks, see James, Witten, Hastie, and Tibshirani (2013).

can be improved upon at each stage of the machine learning process, beginning with labeling the training data (and specifying the size and representation of the sample), and continuing into the feature extraction stage, as well as training a classifier. Methods for improving model accuracy in the latter two stages focus on data science techniques that are not addressed in this paper, such as implementing more sophisticated feature extraction strategies, as well as tuning the parameters of the classifier. However, in reviewing discrepancies of the hand-scorer classifications, some notable patterns emerged that offer promising recommendations for how to improve the labeling stage of the process, and subsequently the model's accuracy.

First, during the hand-scoring training process, emphasis should be placed on the importance of carefully reading the entire piece of student writing. In some cases, only a small portion of the text may be indicative of harm. While the content at the beginning of the text may be benign in nature, it may either quickly escalate into alarming content, or the reported harm may be inserted into the text as a nonsequitur. In such instances, if the hand-scorer does not read the entire text closely, then the text could easily be misclassified. Second, the training materials should include more examples of writing that is both fantastical in nature and containing violent words or phrases. The apparent fictional aspects of the story may dissuade some hand-scorers from taking the content literally and instead assign the text a label of *Normal (Level 1)*, whereas other raters adhered more closely to literal nature of the rubric, assigning the text a *Concerning (Level 2)*. Clear guidance for this type of writing in the training materials can reduce such inconsistencies. Third, the hand-scoring training materials should provide more examples of writing in which students expressed dissatisfaction toward themselves or their lives. It is sometimes difficult to differentiate between writing akin to signs of depression, self-loathing, or anxiety (*Concerning)* and writing that may be more indicative of self-harm (*Alert)*. Providing more examples of such writing can assist in delineating the boundary for hand-scorers.

Limitations

Regarding the rubric development, a number of decisions were made that, if they were different, may have resulted in different criteria for the rubric. For example, the analytic sample used for the qualitative coding exercise consisted of short texts composed of 10 words or less. While this approach afforded a larger quantity of texts to be included in the sample, it was accompanied by the concern that the rubric criteria may be inadequate in guiding hand-scorers to label longer papers. However, in classifying 13,000 pieces of student writing, which included papers of all lengths, there was no evidence that the rubric did not generalize to the subject matter of longer texts (though it is true that the training materials did not provide specific instructions for longer texts). Even still, it is possible that longer papers, not included in this study, may contain topics currently unaccounted for, and an in-depth comparative analysis of shorter papers and longer papers could further inform the generalizability of the results reported here.

Another limiting aspect of the rubric development is that it, in part, relied on the particular framework of coding negative emotions. It is possible that a different analytic framework may have resulted in different rubric criteria. Additionally, the qualitative coding was conducted by a single rater, and while the results were then vetted by a hand-scoring director, there is no inter-rater reliability for the codes.

Regarding the training data used in this study, there are two key limitations. First, the examples of student writing that comprised the training data encompassed all grade levels. On the one hand, this approach leads to a model that represents the vocabulary of all ages. On the other hand, this reduces the number of examples in the training data for each grade level. An alternative approach would be to calibrate grade-specific models, in order to be more sensitive to the lexicon of social and developmental stages of students. Second, the training data over-represent the *Concerning (Level 2)* and *Alert (Level 3)* examples, which are rare occurrences in operational data. An important next step for model validation is to classify a sample of data that more accurately reflects the rarity of crisis papers. This will provide more insight into the model's ability to detect true positives, while ignoring true negatives.

Conclusion

While being mindful of these limitations and revisions to the hand-scoring training procedure, this crisis paper rubric has potential to improve the overall process of detecting crisis papers by alleviating the burden that is placed on hand-scorers who must decide how to handle questionable pieces of text, by guiding their classification decisions. This will lead to consistently labeled data that can be used to train a model. Additionally, an automatic detection system that relies on the design of this three-level rubric will automatically triage student writing, which will assist school personnel in determining which pieces of writing should be prioritized. The ultimate goal is that this rubric can assist in improving the crisis paper detection process, so that students who are in need of intervention can receive appropriate help in a timely manner.

Acknowledgments

We thank Jon Cohen for his support of the project. We appreciate Chris Ormerod's help with modeling, as well as Amanda Adams and the scoring staff at the American Institutes for Research for their scoring work. We would also like to thank Derek Briggs for his feedback on prior iterations of this paper.

References

America's Health Rankings. (2019). *Health of women and children report 2019*. United Health Foundation. Retrieved from https://www.americashealthrankings.org/learn/reports/2019-health-of-women-and-children-report

Brown, B. (2018). *List of core emotions*. Retrieved from https://brenebrown.com/downloads/

Brundin, J. (2019). *Teens under stress*. [Broadcast/Digital]. Colorado Public Radio. Retrieved from https://widgets.cpr.org/teens/index.html

Centers for Disease Control and Prevention. (2019). *Facts about mental disorders in U.S. children*. Retrieved from https://www.cdc.gov/childrensmentalhealth/data.html

Corbin, J., & Strauss, A. (2015). *Basics of qualitative research* (4th ed.). Thousand Oaks, CA: Sage.

Council of Chief State School Officers and Association of Test Publishers. (2010). *Operational best practices for statewide large-scale assessment programs*. Retrieved from http://programs.ccsso.org/projects/operational_best_practices/Operational%20Best%20Practices%20final.pdf

© 2020 by the National Council on Measurement in Education

Dessus, P., Lemaire, B., & Vernier, A. (2000). Free-text assessment in a virtual campus. *Proceedings of 3rd International Conference on Human System Learning (CAPS'3)*. 61–76.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. In *EdMedia+ Innovate Learning* (pp. 939–944). Association for the Advancement of Computing in Education (AACE).

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, *10*(1), 1–309.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Kintsch, E., Steinhart, D., Stahl, G., Matthews, C., & Lamb, R. (2000). Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments*, *8*(2), 87–109.

Raschka, S., & Mirjalili, V. (2019). *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd.

Smarter Balanced Assessment Consortium (2014) *Hand-scoring rules*. Retrieved from http://www.smarterapp.org/documents/Smarter_Balanced_Hand_Scoring_Rules.pdf

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13.