Cambium Assessment
a cambium company

Defining At-Risk Student Responses

Amy Katrina Burkhardt
University of Colorado at Boulder

Susan Lottridge, Sherri Woolf, and Chris Ormerod
Cambium Assessment, Inc.

Paper presented at the annual meeting of the National Council on Measurement in Education (NCME)

April 4 – April 8, 2019, Toronto

Abstract

In this paper, we argue that the current guidelines for flagging "at-risk student responses" (i.e., student writing that is indicative of potential or experienced harm) are inadequate for assisting human graders in consistently identify these pieces of text. The consequences of using inconsistently labeled data to train a statistical classifier to automatically detect these responses can be two-fold. First, student writing is flagged when it is actually normal, which places a strain on the resources allocated for reviewing alerted text. And, even more problematic, students who are asking for help may be overlooked. As such, this paper puts forth an empirically-based and expert-vetted definition that consists of three tiers: (a) normal responses, (b) concerning responses, and (c) alert responses. To demonstrate the practicality of using this definition, we conduct an experiment where six human raters locate 13,000 pieces of student writing within these different categories, and find that, on average, they exhibit good agreement (an exact agreement rate of 87%, adjacent agreement of 11%, and a non-adjacent agreement of 1.4%). We also showcase how neural networks can classify these examples of student writing and find that the agreement statistics between the

machine and adjudicated scores are only slightly lower than the agreement statistics of two human raters. While we are cautiously optimistic about our findings – that our definition can be modeled by both humans and machines – we recommend clarifying particular parts of the rubric for rater training, as well as revisiting the parameters of the neural networks for improving model performance.

Defining At-Risk Student Responses

Overview

In accordance with the best practice guidelines for large-scale statewide assessments, student responses that contain content that "constitutes an immediate and potential threat of harm, violence, abuse or illegal activity" should be flagged for further investigation (CCSSO & ATP, 2013). Historically, flagging such text has been the responsibility of human graders, and the absence of a published definition of at-risk student responses suggests that this task relies, at least partially, on subjective judgement and is treated as an ancillary, ad-hoc part of the scoring process. However, recent technological adaptations within online testing platforms gives cause to revisit and refine existing criteria intended to capture the spirit of this guideline and explore how, and to what extent, this particular type of student writing can be detected by both human graders and machines.

Automated scoring systems are being used more widely in operational settings to score online assessments than ever before. While there is no published study on the use of automated scoring in state assessment programs, technical reports and anecdotal evidence suggests that use is growing and that automated scoring is used in concert with human scoring, whereby a score is assigned by either the engine or the human rater. For example, in the 2016-2017 school year, PARCC assessments serviced six states, the Bureau of Indian Education, and the District of Columbia, and almost 70% of the scores for the prose constructed-response tasks were assigned by an automated essay scoring system (Pearson, 2018). Cambium Assessment, Inc. (CAI), which is the context for this present research study, employs automated scoring within eight state summative testing programs and ten state interim programs. CAI uses machine scoring on about 75% of the essay responses in the summative testing programs, and on all essay responses in the state interim programs. Furthermore, they have recently included automated scoring for short-answer items in five state interim programs.

Concurrent to these advancements in scoring, online testing enables the storage and review of informal writing. Previously, students would be constrained to a text field provided for a given prompt, but now they are sometimes equipped with a digital notepad[1] where, alongside drafting answers to test questions, they can also write down any other thoughts and self-reflections. Though these virtual pieces of scratch paper are not scored, they should arguably still be reviewed, along with the rest of the assessment, to ensure that the student isn't reporting harm or requesting intervention. Although such review would add to the scope of human rating beyond the provision of scores, this influx in student writing can be subjected to automatic detection, by means and methods similar to existing processes associated with an automated scoring system, intended to flag other forms of aberrant student writing (e.g., off-topic responses). Automated scoring systems also offer the ability to detect such responses quickly, with a day or even hours of submission.

While such detection systems rely on automated techniques, the underlying classifiers of these methods depend on a set of labeled

[1]https://assess.smarterbalanced.org/student/Projects/SBAC/Help/help.html

student writing. As such, inconsistently classified student writing in the training data will adversely impact the accuracy of the model, which results in not only unnecessarily identifying students who are not in need of intervention, but also increases the risk of ignoring troubled students. It follows, then, that a clear definition that can be consistently interpreted by humans, needs to be established.

Currently, there is not an agreed-upon definition across state interim and summative assessments. However, the following hand scoring guidelines outlined by the Smart Balanced Assessment Consortium instruct graders to flag the following (but not limited) 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 (n.d., p., 10).

While this list provides salient examples of a student either being harmed, or harming someone else, these descriptions lack examples around what the CCSSO describes as the *potential* for harm, violence, abuse or illegal activities. That is, even if a student doesn't explicitly report a harmful situation, their response might still contain language that gives pause to the human graders, who are hesitant to ignore a veiled request for help or a covert threat. Take, for example, when students use non-literal but violent phrases and hostile language to express frustration over a test, or when a student conveys that they feel that they are inadequate and can't measure up to academic expectations. Students also make use of other emotive language, such as sadness or loneliness, that express negative sentiment towards themselves and their lives and, while hinting at a request for intervention, they don't explicitly report imminent harm. Examples of such text include the hyperbolic and frustrated statement, "kill me now," and this piece of text exhibiting depression (that may or may not be test-related), "I hate my life." And, while the emotion expressed in such responses doesn't configure into the Smarter Balanced examples outlined above, we explore the role of emotion in the types of responses that complicate existing instructions for flagging at-risk student responses. Having a reasoned definition combined with a consistent procedure for the varieties of responses can support a correct and timely intervention of a potentially troubled student.

This present study attempts to clarify the definition of at-risk student responses by proposing a tiered approach for classifying such responses, in which we then delineate between three different types of student writing: a) normal responses; b) responses that are concerning in nature; and, c) responses that adhere to the existing definition of reporting or causing harm to the self or someone else. This design aims to guide human graders in making consistent classifications by removing the polarity of benign vs. problematic responses and by making allowances to ensure that students in need of intervention are not being overlooked. We then explore the consistency of humans when adhering to this tiered definition, as well as the agreement between humans and machine multi-class classification. Figure 1 outlines this high-level process.

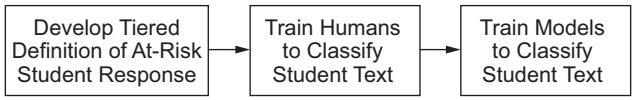

Figure 1. Overview of Project

As such, this paper is guided by three different research questions; the results and data generated from each question informs the following one.

1. **Research Question 1: Defining At-Risk Student Responses**
    a. Given that many of the responses that may be indicative of *potential* harm seem to be emotional in nature, how can the use of an emotive qualitive coding scheme inform and support the development of a tiered definition of alert?
2. **Research Question 2: Humans Detection of At-Risk Student Responses**
    a. Can humans distinguish between the three categories of the tiered definition set forth in the first research question?

i. What are the inter-rater agreement rates of humans?

ii. Which tier(s) were the easiest and which were most challenging for humans to agree on the placement of student writing?

3. **Research Question 3: Machine Detection of At-Risk Student Responses**
   a. Using the labeled training data from the previous research question, how successful are statistical classifiers at distinguishing between these three levels?
      i. How does the performance of a linear classifier compare to neural networks that vary in complexity?
      ii. How do different vectorized representations of student writing impact model performance? That is, how does a bag-of-words approach compare to word embeddings?

We then tie these three components together to discuss affordances and limitations of this approach to defining at-risk student responses, making recommendations for implementing such a model in operational use.

Background

There are only two studies exploring the issue of detecting alerts within online assessments (Ormerod & Harris, 2018; Burkhardt, et al., 2016). One challenge inherent to this research is the rarity of these student responses. To address this problem, "synthetic" student responses data, gathered from publicly available data online, have been used to supplement the expected and typical student writing.

As noted above, this study took place at Cambium Assessment, Inc., as did the work of Ormerod and Harris (2018) who collected data from a web forum called "teen line"[2] that provides a platform for teenagers to share their problems and ask for help from other teenagers. Ormerod and Harris (2018) also used data from reddit, a discussion website where anonymous users can publicly share and discuss questions, ideas and links with others. Within the reddit website, discussion themes are separated into different forums, which are known as subreddits. The synthetic data from both sources were reviewed by human graders to ensure that the data reflected the tone, content and writing style consistent with student responses.

Burkhardt et al. (2016), also supplemented normal student responses with text taken from various subreddits that focused on symptoms associated with child abuse and neglect, such as self-harm and suicide (as well as explicit reports of abuse). While Burkhardt et al. (2016) relied on an underspecified definition to flag any response that exhibited symptoms of a student in need, Ormerod and Harris (2018) used a definition similar to that of Smarter Balanced and noted that there wasn't a systematic way to classified potential harm (and that responses of potential harm were sometimes classified as alerts).

While both papers explored the model performance of different classifiers, Burkhardt et al. (2016) focused on blending results from support vector machines, gradient boosted regression trees and random forests, along with a rule-based model. The best ensembled model, consisting of the voting rule that a response was flagged as an alert when two or more of the four models agreed that a response was an alert, produced a recall of .92 and precision of .84 on a test set.

Ormerod and Harris (2016) compared a baseline logistic regression model to 16 neural networks varying in complexity. Instead of a binary classification, the output from each model was the confidence that a piece of student writing was a true alert. The percentage of the alerts caught by each model was evaluated based on the different thresholds of responses being able to be reviewed, ranging from .1% to 4%. Not surprisingly, they found that the largest percentage of alerts caught by the model (98.7% of all alerts present in the test set) occurred when 4% of all test responses were being reviewed.

[2]https://teenlineonline.org/board/

Adjacent research in efforts to detect internet bullying, toxic comments, and indicators of poor mental health in text also informs this present study. Facebook recently announced their efforts in detecting suicidal language within users' posts[3]. While the manual reporting of posts has been ongoing for some time (where users can then be connected to support resources), they have moved on to automatic classification, and have faced similar challenges in this present study. First, they noted that words that are oftentimes associated with suicide are also benign in other contexts – words, such as "kill", and "die" and "goodbye." To exemplify this issue, they provided an example similar to what we have discussed above: "I have so much homework I want to kill myself." And, while Facebook determined that this response does not meet the criteria to be flagged, it does introduce the complexities of this classification task, and the reliance on human judgement required in determining the subtle lexical cut-offs for what should and should not be flagged.

Another problem Facebook encountered was developing the labeled training set. For the examples of suicidal text, they relied on posts that had been flagged by users and were manually vetted to be truly suicidal. They then combined these positive examples with negative examples – not of normal posts, but rather of posts that had been flagged as potential suicide posts but were determined not to be a person at risk of suicide. In creating such a training sample, they suggest that it will help develop a "more nuanced" understanding of what a suicidal pattern is and what it isn't. Various sequences of words (n-grams) were used as features to train a random forest classifier. The classification of each post is then subject to review by Facebook personnel, and if deemed necessary, the post's author is then provided with resources, or escalated to local authorities.

Along the lines of suicide prevention, Alhanai, Ghassemi, and Glass (2018) demonstrated that they could model the question-answer portion of a depression screening test conducted by a virtual agent and a human subject. They used 142 transcriptions from the interactions, as well as from 170 possible questions (e.g., "How are you?", "Do you consider yourself to be an introvert?"), which are stored in a publicly available corpus called the Distress Analysis and Interview Corpus (DIAC). They modeled the binary variable of depressed or not, as well as the multi-class classification of the severity of the depression. The text features that they used were then transformed into word embeddings (which are vector space representations of words aimed at mapping semantic meaning into geometric space) and used as features to train a number of different models, ranging from a baseline model of logistic regression, to more complex neural networks. They found that their models outperformed the baseline approaches that had previously been conducted, using the same corpus.

While not specific to harm, or symptoms associated with potential self-harm, derogatory comments and bullying may also be included in student writing about past, present or potential harm. Google has created a tool that intends to detect toxic comments online – they define such text as a "rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion." The classification tool is called Perspective, and the labeled training data was collected by presenting internet comments to raters and asking them to rate each comment on a scale from "Very Toxic" to "Very Healthy"[4].

Nobata, Tetreault, Achint, Mehdad, & Chang (2016) used a regression model to identify abusive language in online user content. They document four classes of features: n-gram features (i.e., 3-5-character n-grams and unigrams and bigrams), linguistic features (e.g., length of comment, average length of word, number of punctuations, number of politeness words), syntactic features (which uses a dependency parser to capture long

[3]https://newsroom.fb.com/news/2018/09/inside-feed-suicide-prevention-and-ai/

[4]https://www.perspectiveapi.com/#/

range dependencies between words), and distributional semantics (e.g., pre-trained word embeddings). They found that the word embedding features were powerful when combined with standard NLP features, although character n-grams also worked well.

The data used in Nobata, Tetreault, Achint, Mehdad, & Chang (2016) was also used in prior research (Djuric, et al., 2015) which used learning techniques of neural language modeling. The data in both of these studies were sampled from comments posted on Yahoo! Finance and News during the period between October 2012 and January 2014 and were labeled by human raters as either "clean" or "abusive" text. The research in this present paper also uses internet comments to build the portion of the training sample considered disturbing content.

This present study is situated in the existing research in the following ways. First, it sets forth a clear and detailed three-tiered definition that is empirically based and vetted by educational hand scoring experts. To defend the practicality of using this rubric we provide results on how well humans could distinguish between different tiers of true student writing (using a sample comprised entirely of student writing, without the use of synthetic data). Finally, we demonstrate the capability of various machine learning approaches in modeling human labels.

Defining At-Risk Student Responses

Methods

In order to define at-risk student responses, we applied the tiered definition framework to the context of large-scale handscoring efforts. The final structure of the tiered definition is as follows: Within each of the three tiers, we present a number of categories that further classify and characterized a piece of student writing. These categories and their supporting details are intended to guide the human grader in locating precisely where a given piece of student writing belongs within the definition.

The categories were developed, in part, through the use 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 phrase, represented a characteristic or phenomenon of interest in the data (Corbin & Strauss, 2015). These codes were then further aggregated to generate the categories within each tier and were continuously vetted by experts for appropriateness within the context of operational use. Figure 2 presents the overall process of iteratively combining qualitative coding with expert oversight with the end product as the three-tiered definition of at-risk student responses. While only a single person conducted the qualitative coding that was used to engage in a systematic process of reviewing and summarizing responses to build out a definition, the iterative process of mapping codes to tiers was done through the oversight of an expert staff member in hand-scoring.

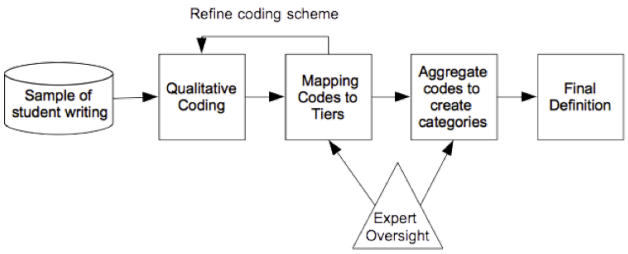

Figure 2. Qualitative Coding Approach for Defining At-Risk Student Responses

Data

For the qualitative coding, we focused only on responses that were 10 words or less in length; a restriction that allowed for a greater number of student responses to be reviewed. 8,344 responses were included in this analytic sample. 7,791 responses been previously classified as an at-risk student response by a neural network model but were verified by human raters to be normal (these false positive responses are referred to as "not-alert" responses). Responses that comprised this sample were student responses of any grade and content area from any state served by CAI.

In addition to the "non-alert" responses, 553 of these short responses were those that both humans and the neural network classified as true at-risk student responses (referred to as "alert" responses). We expected that these student responses would all be responses that are consisting of language indicative of harm, in alignment with the only existing definition Smarter Balanced definition outlined above, and one goal of coding these was to develop a coding scheme to capture harm-related student writing.

Developing the Qualitative Coding Scheme

We hypothesized that student writing that does not fall cleanly into the binary classification of at-risk student response vs. normal response are those that are often characterized as being emotive in nature. As such, through the qualitative coding process, we explored the role that emotion plays in student responses, and how codifying these emotions might help us define at-risk student responses. Therefore, to characterize the content of this sample of responses, our qualitative coding scheme was comprised of the following three types of codes: (1) codes characterizing negative emotions, (2) codes capturing reports of imminent, past, or current harm, and (3) codes describing responses that do not exhibit a clear instance of emotion. These three groups of codes helped us understand the following about the text within our analytic sample: the proportion of student writing that could be codified as expressing emotion, the types of harmful situations and events students are reporting, and the non-emotion codes that can characterize the remaining student responses. We first coded "non-alerts" and then proceeded with the "alert" sample of student writing.

Coding of Non-Alerts

We started the iterative process by applying codes to student writing with a list of nineteen negative emotions: anger, anxiety, blame, disappointment, disgust, embarrassment, fear, frustration, grief, guilt, humiliation, hurt, jealousy, loneliness, overwhelmed, regret, sadness, shame, and worry (Brown, 2018). For each piece of student writing, we first determined whether or not the student expressed emotion. If so, then the piece of writing was assigned one of the emotion codes. If no emotion was present in the response, then we took a more inductive approach to coding and created a non-emotion code that accurately characterized the text. Some of these codes were: physical discomfort, language issues, reports of not having opportunity to learn the material, and a catch-all category for written work that couldn't be categorized by any of the emotive or non-emotive codes (most often, these responses were responses that were on-topic for a particular test item and were sometimes flagged for language that, without context, might seem similar in nature to at-risk student responses). In the event that the classification of a student response was unclear, it was coded as "other." This initial coding phase was followed by a series of iterations to review codes that were identified as "other." Additional codes were developed to classify these pieces of student text when it made sense to do so, and when it was difficult to differentiate between two codes, we would collapse these codes into a single code.

Approximately half of the responses fell into the catch-all category for written work that couldn't be categorized by any of the emotive or non-emotive codes. Of the remaining pieces of student writing, 70% of them were characterized by an emotion code. Of the original nineteen emotion codes, eleven 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 (e.g., a student reports that they are emotionally exhausted and can't 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., negatively compares self's actions with self's standards). Frequent non-emotion codes were those where the student simply stated that they didn't know the answer (without exhibiting emotion) and reporting complaints of physical discomfort.

Coding of Alerts

To gather information on types of harmful situations and events students reported, we engaged in the following coding process: We first reviewed each of the alert responses to determine if the response was explicitly stating a past, present or imminent harmful situation. If it was, then a harm code was applied through an inductive approach. Otherwise, the response underwent the same procedure outlined above for non-alerts (assigning either an emotion or non-emotion code to each piece of student writing).

While the primary purpose of the coding was to gather empirical evidence of the types of harm that students are reporting, we concurrently found that our coding results were contrary to our hypothesis that all of these "true alerts" would be coded as imminent, present or past harm. The emotions of frustrated, overwhelmed, anger, shame, bored and sad comprised 60% of this sample of alerts. As such, this coding process illuminated a possible artifact of the "status quo" flagging approach, absent of a clear and non-binary definition.

Mapping Codes to Tiers and Human Expert Review

These codes continued to be refined as they underwent the mapping process which we linked each qualitative code to one of the three tiers. For example, the code Frustrated was further refined into two sub-codes: Frustrated: Hyperbolic (*Frustrated statements that include violent language*) which was mapped to Tier B, and Frustrated: Irritated (*Characterized by being annoyed or irritated*) which was mapped to Tier C.

As shown in Figure 2, This mapping process was performed in-step with a handscoring expert, who had 25 years of experience of managing and directing hand scoring centers tasks with reviewing and assigning scores to millions of pieces of student writing. The alignment between the codes and the tiers was thoroughly reviewed by expert judgement to ensure that these categories wouldn't be problematic in operational use.

These codes were then aggregated to create categories within each tier. For example, the emotions of sad, lonely, hurt, grief, anxiety and fear were used to describe a sub-category in the Concerning tier: *Signs of depression, self-loathing, or anxiety.*

Results

The final coding scheme is presented in Table 1, which also represents the mapping of each qualitative code to each of the three tiers. The final definition is presented below in Table 2 and was used to train the human raters on how to differentiate between the three types of responses. Note that the Tier A Categories closely reflect the Smarter Balanced definition of an alert, whereas Tier B delineates cases of student writing that exhibit potential harm. The categories of the Tier C responses outline those type of responses that are similar to Tier B responses but have been deemed to not meet the criteria of either harm or potential harm.

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

Qualitative Code	Description
Tier A: Alerts*	
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

Qualitative Code	Description
Tier B: Concerning	
Overwhelmed	Request for help that isn't 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**	Non-threating mention of a gun
Suggestive**	Text suspect of student being the perpetrator or victim of violence
Sexual **	Sexual imagery without threatening abuse or reporting abuse
Tier C: Normal	
Frustrated: Irritate	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 **	Test-related request for help
OTL**	Reports not having opportunity to learn
IDK**	Reports not knowing content
Technical**	Technical issues with computer and test
Other	Normal responses that do not fall into any of the above categories

* The only emotive code for alerts is identifying cases of extreme depression (Sad:Extreme)

** Non-emotive codes

Table 2. At-Risk Student Response Definition

Tier	Category	Details
A - 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.

Tier	Category	Details
B - 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; non-threatening 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 towards self. Includes hyperbolic language about wanting to die.
C - Normal		
	Hating another entity.	Hate towards 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 note characterized by any of the criteria outlined in this definition.

Human Detection of At-Risk Student Responses

The second portion of this study was to explore how well human graders can differentiate between the three tiers of the definition. For this exploration, we trained six human raters to review thousands of student responses and locate each response within one of the three tiers by assigning a label of A, B or C. All pieces of student writing were labeled by two raters. Responses receiving discrepant labels were sent for adjudication by an expert rater.

Methods

Data. 13,000 responses were used in this study. Five-thousand responses were identified as containing disturbing content by human raters (using the Smarter Balanced guidelines), and the other 8,000 were deemed normal by human raters. Although 8,000 were considered normal, we ran these responses through a series of trained neural networks and selected the responses that had the highest probability of being an alert response. The purpose of this approach was to capture a sample of student responses that might be difficult to classify in the context of a binary definition. In other words, we were interested in responses that contained characteristics that might blur the lines between an alert and non-alert – i.e., Tier B responses. For this sample of responses, there was no length restriction, and consisted of writing across all available grades and content areas.

Training and Rating Procedure. Six human raters were hired for this task from a pool of raters that have previously scored open-ended assessment items for CAI, but none had prior experience classifying responses into three tiers of alerts. However, at least one had experience with determining the student authenticity of the “teen line” responses used in Ormerod and Harris (2016). While these raters had experience with scoring essays, none were experts in these types of alerts (though some may have encountered them during routine grading assignments). This study requirement allowed us to examine our hypothesis that locating student writing within the three different tiers is a trainable phenomenon.

Upon the development of the rubric, we created training resources and trained the six humans to label 12,917 responses according to

the definition (83 responses were flagged by the humans as being written in Spanish and were removed from the sample). Upon the review of the rubric, we administered three qualifying sets to the raters, each containing ten pieces of student writing. The raters were instructed to assign one of three labels to each piece of student writing to denote which tier the writing belongs to. The results from these sets were sufficient to qualify each rater to proceed in the process (see Table 3).

This scoring consistency prompted the next portion of the training, which was for all raters to be assigned a computer and begin classifying the 12,917 responses. On the screen, one piece of student writing was presented at a time, and the rater determined which tier the student response belonged to. For the first two hours, the raters were able to confer with one another prior to selecting the tier. After two hours, a set of three pieces of student writing (i.e., validity papers) were presented to the raters, to ensure that decisions that they were making were still aligned with the rubric. All raters agreed on the three answers.

After the validity check, raters began to score without consulting one another; asking the lead rater questions as necessary. They continued to classify student writing for six days, while agreement rates were monitored throughout the process, and validity papers were administered daily to ensure that human raters continued to be consistent with the rubric. Agreement rates and score point distributions were monitored for two key issues: 1) to ensure that agreement rates were high, and no intervention was needed to clarify the rubric; and, 2) to detect any characteristics of the data that could indicate that raters were going "off-script" from the rubric and making changes that might be different from the rubric (which would result in even higher agreement rates). As such, no intervention was needed, as the agreement rates and score point distributions remained steady over the course of the six days. All responses were double-labeled and underwent expert review when scores were discrepant.

Results

Table 3 presents the results of the three qualifying sets; no rater mislabeled more than two pieces of student writing, and the modal score was 90%. These results were sufficient to qualify each of the raters.

Table 3. Percent correct for three qualifying sets of 10.

Qualifying Set	Rater 1	Rater 2	Rater 3
1	90	100	90
2	90	90	80
3	100	90	90
Qualifying Set	**Rater 4**	**Rater 5**	**Rater 6**
1	100	90	100
2	90	80	80
3	90	80	90

Table 4 presents the cross-tabulation between scores assigned by rater 1 and rater 2, using rater 1 as the basis for comparison, the category proportion agreements are included parenthetically.

Table 4. Cross Tabulation between Rater 1 and Rater 2

		Rater 2			
		Alert	Concerning	Normal	Total
Rater 1	Alert	2391 (83%)	401 (14%)	83 (3%)	2875
	Concerning	464 (21%)	1413 (65%)	289 (13%)	2166
	Normal	102 (1%)	317 (4%)	7457 (95%)	7876
	Total	2957	2131	7829	12917

The exact agreement between the two raters was 87%, non-adjacent agreement was 11%, and non-adjacent agreement rate was 1.4%. The "normal" written responses were the easiest for readers to agree on (exact agreement rate of 95%), followed by alerts where raters agreed on 83% of the placements of pieces of student writing, and lastly the "concerning" results were the most difficult for agreement (exact agreement of 65%).

Some notable patterns emerged from an inspection of text of the non-adjacent ratings (where one rater identified the response as "normal" and the other rater identified it as an "alert"). At times, the discrepancy seemed to be due to an accidental misclassification of "normal." In such cases, the text was long and irregularly formatted (e.g., a lot of carriage returns). The content at the beginning of the text was benign in nature, but it either quickly escalated into alarming content, or a student reported harm within a small portion of the text. If the rater didn't read the entire text closely, it would be easy to accidentally flag such response as normal.

Another characteristics of these discrepancies were responses that seemed to be written to be both fantastical in nature as well as violent. The apparent fictional aspects of the story may have dissuaded some raters from taking the content literally and assigned a "normal" whereas another rater adhered more closely to literal tendency of the rubric.

Some of the responses that seemed to cause a lot of adjacent disagreement between "concerning" and "alerts" were when the student expressed dissatisfaction towards themselves or their lives. Some of these pieces of student writing were hard to differentiate between self-loathing (e.g., "teenage angst") and something more extreme and indicative of potential self-harm.

As noted above, the original composition of these sample consisted of approximately 5,000 pieces of student writing that were previously flagged as alerts under the Smarter Balanced definition, and 8,000 pieces of text that were considered normal (40% and 60% respectively). Under the new tiered definition, and the final, adjudicated score, we note that the new distribution is the following: 23% were flagged as alerts, 17% were flagged as concerning, and 61% were flagged as normal. While most of the normal responses remained as such (with 4% shifting to "concerning" and 1% shifting to "alert"), we see a larger redistribution of responses within the alert sample; of the 5,000 responses, only 57% of them remain as alerts, 36% are now considered "concerning" and 6% were now identified as normal.

Machine Detection of At-Risk Student Responses

Methods

This final component of the study explores the performance of statistical classifiers in distinguishing between different levels of student responses (normal, concerning, or alerts). We compared five different models, which allowed us to observe the stability of modeling and to see whether -- and how -- complexity adds value. Through this model building process, we compared different inputs and classifiers. We compared inputs of fixed length vectors representing each text document using Latent Semantic Analysis (LSA) to sequences of word embeddings. We also compared a linear classifier to neural networks to test to see if the non-linearity of neural networks leads to improved model performance. Figure 3 presents the five models, their associated characteristics and inputs, an overview of the classifiers, LSA, and word embeddings are presented below.

In order to create the LSA matrix and training the linear support vector classifier, we used scikit-learn's implementation (Pedregosa et. al., 2011). When building neural networks, we used the deep-learning framework for Keras (Chollet, 2018) as the front-end, which relies on Tensorflow as the backend engine (Abadi, et al., 2015). Word embeddings were trained using a Word2Vec model and the Python package Gensim (Rehurek and Sojka, 2010).

To evaluate the different models, we computed agreement statistics between an adjudicated human grader and the machine score, namely exact, adjacent, and non-adjacent agreement, as well as quadratic weighted kappa QWK). The score point distributions across the three categories was also computed

and compared between the machine and the adjudicated human score.

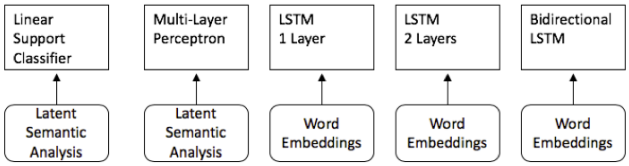

Figure 3. Input and Classifiers for Five Models

Data

Of the 12,917 responses scored by humans, we removed a small number of responses that were uncharacteristically long to ensure that they wouldn't be an issue in training. Training and validation were conducted on the remaining 12,901 responses. All of the competing models were trained on the same training sample and validated on the same test sample. We trained each classifier using the adjudicated human score from the human-scoring portion of this experiment.

Features

Latent Semantic Analysis (LSA)

The input for the first two models, the linear support vector classifier and the feed forward neural network (i.e., the multi-layer perceptron) relied on a dimension reduction technique called latent semantic analysis (LSA), which is a variance decomposition technique that captures the similarities of words and pieces of writing (Landauer and Dumais, 1997). LSA has been used in the context of automated scoring (Foltz et al., 1999) and other intelligent tutoring systems (e.g., Kintsch et al, 2000; Dessus et al., 2000), and is thought to reflect the salient semantic information of a text document.

The process to transform the student responses into a matrix was as follows: First the responses underwent a small amount of pre-processing to remove html tags. We then transformed each piece of student writing into vector representations by first counting the number of unique unigrams (single words) within each response in the training sample. We then applied a normalization approach called term frequency-inverse document frequency (tf-idf) to these vectors, so that n-grams that occur frequently across all pieces of student writing in the training sample are weighted less heavily compared to more unique words. This matrix then underwent the dimension reduction technique, which is specified as support vector decomposition, or SVD, in scikit-learn. In performing this reduction, the number of dimensions of the normalized vectors was then reduced from 37,000 to an optimal 3,500 components. This number was selected through a small bootstrap study undertaken to pick the optimal number of components.

Word Embeddings

Word embeddings are numeric representations of words in low dimension vectors (in this study, 200-dimensional space) that capture semantic relations among words and are oftentimes used in a sequence labeling task. Word embeddings are impacted by the corpus on which they are trained, and so, instead of using pre-trained word embeddings, we trained our own embeddings on a corpus of student writing consisting of 1.12 million words.

Classifiers

Support Vector Machines

Support vector machines (SVMs) have been shown to perform well on a variety of classification tasks (James, Witten, Hastie, & Tibshirani, 2013). In an m-dimensional vector space, an (m-1)-dimensional hyperplane is fit in such a way that the distance between the hyperplane and the nearest training feature vectors from each of the classes (the "support vectors") is maximized. New documents are then mapped to the same space, and the classification is determined based on which side of the hyperplane the document is mapped to. Multi-classification was obtained through a one-vs-the-rest approach, which means that the model was fit three times, each time comparing one of the classes to the other two classes. Default parameters were used in scikit-learn, and so a cross-validation framework was not required to pick hyperparameters.

Feed-Forward Network

A multi-layered perceptron, or a standard feed-forward neural network was used to assess whether non-linearity leads to improved performance, while still using the fixed-length inputs of an LSA matrix. To describe the basic framework of a neural network model, it can be helpful to begin with a logistic regression model, which can be written as the following:

$$LogReg(\boldsymbol{x}) = g(\boldsymbol{xW} + \boldsymbol{b})$$

Where $\boldsymbol{W}$ is the weight matrix, $\boldsymbol{x}$ are the features, $\boldsymbol{b}$ is the bias term (or intercept), and g is the sigmoid function. The pivot, then, from this representation to a feed forward network is rather straight forward, and can be thought of as including a hidden layer, which can then be written as this this:

$$NN_{MLP_1}(\boldsymbol{x}) = g(\boldsymbol{xW^1} + \boldsymbol{b^1})\boldsymbol{W^2} + \boldsymbol{b^2}$$

$$\boldsymbol{x} \in \mathbb{R}^{d_{in}}, \boldsymbol{W^1} \in \mathbb{R}^{d_{in} \times d_1}, \boldsymbol{b^1} \in \mathbb{R}^{d_1}, \boldsymbol{W^2} \in \mathbb{R}^{d_1 \times d_2}, \boldsymbol{b^2} \in \mathbb{R}^{d_2}$$

The single hidden layer here is $g(\boldsymbol{xW^1} + \boldsymbol{b^1})$. If we had two hidden layers, each layer can be depicted using the intermediate variables $\boldsymbol{h^1}$, $\boldsymbol{h^2}$ and $\boldsymbol{y}$ (here, the bias vectors are forced to zero in the last layer):

$$NN_{MLP_2}(\boldsymbol{x}) = \boldsymbol{y}$$
$$\boldsymbol{h^1} = g^1(\boldsymbol{xW^1} + \boldsymbol{b^1})$$
$$\boldsymbol{h^2} = g^2(\boldsymbol{xW^2} + \boldsymbol{b^2})$$
$$\boldsymbol{y} = \boldsymbol{h^2 W^3}$$

The vector that results from each linear transformation is a layer (the outer most is the output layer, and all others are the hidden layers). Each hidden layer is followed by a non-linear activation. For logistic regression, g represents the sigmoid function, but in this context, g refers to the activation function, and is not limited to a sigmoid function, but any number of non-linear functions -- some of the most popular being the sigmoid, tanh, hard tanh, and the rectified linear unit (ReLU). For a more in-depth primer on the use of such networks (along with more complex models) in the context of natural language processing, see Goldberg (2017).

For our present study, we trialed an architecture using three hidden layers, each with a different number of hidden units (1000, 250, 50). The output layer is a distribution across all three class assignments, and sums to 1. Between each layer, we applied a regularization technique called *dropout training* to prevent overfitting by preventing the networking from learning to rely on specific weights. In this feedforward network, the activation function applied to each hidden layer is a hyperbolic-tan function, which transforms the values of $\boldsymbol{x}$ into the range [-1, 1]. The architecture of this model was configured based prior findings with similar data.

The training process consists of optimizing a loss function as a function of the parameters defining the network over all the training data for a fixed number of times. Each iteration over the training data is called an epoch. We ran the training over as many epochs as it took for the associated loss function to not improve within some tolerance. We visually determined that the loss function showed little improvement after 10 epochs. This was determined by first dividing the training sample into a training and validation sample, and then observing that the training and validation accuracy and loss both improved up to, and including, 10 epochs, and then diverged after 10 epochs. To minimize the error in the training data, we applied the loss function of categorical cross-entropy.

Recurring Neural Networks and LSTMs

Recurrent neural networks (RNNs) can be used to model sequences of words, whereby the next word in a sequence is modeled on the previous words. Very simply put, these are networks with loops in them, which allows information to be passed from one step in the network the next step. While RNNS are often times used for language generation tasks, such as machine translation, we can also use them to output a class label at the last sequence element.

A distinguishing feature of such networks is that they can be used to analyze inputs of

different lengths (this limitation was addressed in our feed-forward network's by using a fixed-length input of 3,500 features). Furthermore, with recurrent networks, features that explicitly depend on word order can be learned (whereas semantic meaning across words was not retained in the LSA "bag-of-words" approach). This makes recurrent neural networks ubiquitous in natural language processing tasks, and a promising approach for improving the classification accuracy this present study.

However, learning sequences where there are dependencies between disparate elements in the sequences -- such as long pieces of student writing where the most recent words might not provide all of the necessary context to predict the appropriate tier classification -- presents a major problem. In a simple neural network with traditional optimizers based on the stochastic gradient descent methods, the chain rule dictates that the dependence on previous inputs drops off exponentially with length. This problem for recurrent networks is known as the "vanishing gradient problem," and Hochreiter and Schmidhuber (1997) proposed a solution called "Long-Short-Term Memory" (LSTM). Whereas simple RNNs have a structure of a single neural network layer, LSTMs have four layers that interact as a gating mechanism to transfer states of previous units in a way that better learns long-term dependencies. The structure of LSTM-based recurrent networks is further discussed in Olah (2015).

Through the work of language translation tasks, it was noticed that the performance of the neural network performed better if one of the languages was inputted into a recurrent layer of LSTM units in the opposite order to the other language. The solution proposed was that one constructs a recurrent network in which, given a sequence, half the units in a layer are fed the sequence in the correct order while the other half are fed the input in reverse. The resulting construction is known as a bidirectional recurrent neural network, which we implement as our most complex model.

The model we choose to represent the simplest viable recurrent neural network, which we called Model 3, consisted of a single layer of 512 LSTM units. Given any sequence of words, the output of this layer was taken to be the output of the last recurrent units, hence, is a fixed vector of length 512. Linear combinations of the elements of this vector are formed to obtain a single vector of length 3, which is a single value for each tier alert, at which point we apply a softmax function to obtain probabilities that neural network the input. Similarly, Model 4 is defined by stacking 2 layers of LSTM units in the above described manner, each layer consisting of 256 units so that the output of the stacked LSTM layers is a vector of length 256. This output is then sent to a vector of length three where a softmax function is applied. Lastly, Model 5 consisted of 2 layers of bidirectional LSTM units where in each layer we have 128 units fed the sequence in the correct order and another 128 units fed the sequence in reverse. The output is a vector of length 256 from which we obtain a vector of length 3 in the same manner as Model 4.

Results

Table 5 presents the results from the five models, as well as the agreement statistics between the human raters. Though the linear support vector classifier performed similarly to the neural networks, there is a bit of an improvement observed as we move from the SVC to the feed forward network, and slightly more improvement is observed in the two most complex models. The single LSTM layer performed the worst of all the models. However, even the more complex models don't result in large gains for QWK and non-adjacent agreement rates. Additionally, while the score point distributions are consistent with adjudicated human scores, the best-performing model has a higher non-adjacent rate (1.7% higher,) and lower QWK (.04) compared to the human raters. A further examination of the non-adjacent responses reviewed that there were slightly more "normal" responses that were classified as "alerts" by the model (39

responses), compared to the other way around: 31 responses were considered true alerts by the human, but the machine classified them as normal. Of these 31 responses, 10 of them received a score adjudication, noting that they caused some amount of confusion for human raters as well.

Table 5. Validation results from competing models (n = 2584)

	Human	Model 1 LSA + SVC	Model 2 LSA + Feed Forward Network	Model 3 Word Embedding + LSTM (1 layer)	Model 4 Word Embedding + LSTM (2 layers)	Model 5 Word Embedding + Bidirectional LSTM (2 layers)
SPD (%)						
Alert	23	24	23	25	23	24
Concerning	16	11	16.5	17	14	15
Normal	61	64	61	57	62	61
Agreement (%)						
Exact	88	86	86	82	86	87
Adjacent	11	10.5	11	15	12	10
Non-adjacent	1.0	3.5	3.0	3.6	2.6	2.7
QWK	.89	.83	.84	.80	.84	.85

Note: Human agreement rate is computed on the two human raters, whereas Human SPD is the resolved, final score. The agreement rate for the automated engines is computed on the engine score and the resolved, final score.

Discussion and Limitations

The exercise of qualitative coding of emotions, along with a small number of non-emotions codes, allowed us to group and synthesize responses to ensure that our definition was comprehensive of student responses that were either inconsistently classified by humans or misclassified by neural networks. While emotions didn't contribute largely to the Alert tier of the definition, characterizing different emotional responses aided in the construction of the Concerning and Normal tiers. This analysis also exposed that current guidelines may be problematic in consistent classifications: we found that it wasn't unusual for responses that didn't explicitly state harm to be flagged as such by humans.

We do note a number of limitations with this process. The sample used for the qualitative coding exercise was not only a non-random sample, but it also consisted of very brief responses (10 words or less). It is possible that we could have arrived at different results with longer responses. Furthermore, we might expect different results if we used a different analytic framework, other than emotions, to guide our coding work. Additionally, the qualitative coding was conducted by a single rater – and while the results were then vetted for their usefulness in developing a definition by a hand-scoring expert – we have no knowledge of inter-rater reliability for the codes.

As we set forth the proposed definition of this paper, we acknowledge that not all the criteria in the definition are set in stone, nor does our definition necessarily encompass all instances of Alert, Normal, and Concerning responses. We anticipate that additions and modifications can be added to this, based on new data and/or changes in demands from policies or stakeholders, such as states or school districts.

Even though the tiered definition may not be in its final state, the results from our hand-scoring portion of the project does suggest that such a definition is a viable approach in an operational setting. We found that we can be cautiously optimistic with humans being able to differentiate between the tiers, with an exact agreement rate of 87%, adjacent rate of 11% and non-adjacent of 1.4%. The "normal" responses were the easiest for the raters to agree upon, followed by the "alerts." The "concerning" pieces of student writing were the most difficult to agree upon, and possible emphasis in training on differentiating between "concerning" and "alerts" and "concerning" and "normal" might improve rater agreement. One clear area to encourage these efforts would be in further outlining the differences between "extreme depression" from the self-loathing and less severe forms of sadness, and by providing more examples of each.

Though these agreement statistics may be considered a success in the context of hand-scoring test items, the possible implications of misclassifying a student asking for help demands stricter standards. At this point in the research, there is not only room for improvement, but also promising avenues to improve the exact agreement rate, while reducing the non-adjacent rate.

Improvements can be made in both clarifying the rubric and focusing on particular characteristics of writing during rater training, such as more pointedly addressing how to account for fantastical writing, emphasizing the need to carefully read seemingly normal prose, and by providing more examples of the edge cases within the Concerning Tier. Additionally, further research or implementation of this rubric could include a range-finding committee.

When we used these human classifications as training data for our series of models, we found that the best performing model was the bi-directional LSTM model. Although the exact agreement rate is quite similar to the human raters (as is the score-point distribution) the non-adjacent agreement was slightly lower. And, while the difference may be less of a concern in the context of most automated scoring tasks, the implications of a 2.7% non-adjacent rate should give us a bit more pause in this context. Furthermore, given that machines are trained on, and evaluated against, the best available human score, one might expect them to out-perform human raters. On a positive note, over half of these non-adjacent responses were false-positives. Though such mis-classifications have the potential to be a strain on human resources who must review all flagged responses, it is encouraging to acknowledge that not all 2.7% of non-adjacent were instances where a student's request for help would be overlooked. We hope that with a bit more training of human raters, and possibly more data for training the classifiers, we might be able to become closer to human performance.

Addressing other limitations of this study might also improve model performance, such as grounding the parameter choices (e.g., number of layers, number of hidden units in layers, choice of layer) for the neural networks using research-based theory; however, the field is too nascent yet to rely on this type of knowledge, particularly for text. We should also take steps to explain or uncover why discrepancies occurred, as this may help us to refine models. As model complexity grows, the mapping of the modeling result to the response text becomes more challenging. As noted earlier, discrepancies are likely due to characteristics of the response that make it 'borderline,' rarity of the phrasing, or the placement of alert text in the response relative to the length of the response or other words used in the response. Also, classes were intentionally sampled in a way that were somewhat balanced. Training and evaluating performance using imbalanced data that better reflects the rarity of such pieces of student writing would provide more insight into how such filter will do operationally. Additionally, the size of the dataset was relatively small (about 13,000) for modeling.

While being mindful of these limitations and revisions, our findings suggest that the operational use of this definition can be useful

in a number of different ways. First, in contexts where human graders are scoring responses, this tiered definition can reduce the burden that is placed on human graders to decide how to handle questionable pieces of text, by guiding them in their classification decisions. This, in turn, will lead to less noisy training data that can serve to train an accurate and automatic detection system. Furthermore, this definition can offer more flexibility to schools, districts and states as to what types of responses should be further vetted by school personnel. Ultimately, the work in improving the classification of both human raters and statistical models is important so that students who need intervention can receive appropriate intervention in a timely manner.

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C. & Ghemawat, S. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. *Software available from tensorflow.org, 1*(2).

Al Hanai, T., Ghassemi, M., & Glass, J. (2018). Detecting Depression with Audio/Text Sequence Modeling of Interviews. In *Proc. Interspeech* (pp. 1716-1720).

Burkhardt, A., C. Morales, S. Lottridge, S., & S. Wood. (2016). The Automated Detection of Disturbing Content within Online Assessments. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), San Antonio, TX.

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

Chollet, F. (2018). Keras: The python deep learning library.

Corbin, J., & Strauss, A. (2015). *Basics of Qualitative Research* (Fourth ed.). Thousand Oaks: Sage.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015, May). Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web* (pp. 29-30). ACM.

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).

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016, April). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web* (pp. 145-153). International World Wide Web Conferences Steering Committee.

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

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

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

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735-1780.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York: Springer.

Olah, C. (2015). Understanding lstm networks. Retrieved from *http://colah.github.io/posts/2015-08-Understanding-LSTMs*.

Ormerod, C. M., & Harris, A. E. (2018). Neural network approach to classifying alarming student responses to online assessment. arXiv preprint *arXiv:1809.08899*.

Pearson. (2018). PARCC: Final technical report for 2017 administration. Retrieved from: https://parcc-assessment.org/wp-content/uploads/2018/03/PARCC-2017-Technical-Report-Final-03162018.pdf

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research, 12*(Oct), 2825-2830.

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.