SMU Data Science Review

Volume 6
Number 2 *Summer 2022*

Article 11

2022

Using Natural Language Processing to Increase Modularity and Interpretability of Automated Essay Evaluation and Student Feedback

Chris Roche
Southern Methodist University, cmroche@gmail.com

Nathan Deinlein
Southern Methodist University, ndeinlein@mail.smu.edu

Darryl Dawkins
Southern Methodist University, ddawkins@mail.smu.edu

Faizan Javed
Southern Methodist University, faizan.javed@gmail.com

Using Natural Language Processing to Increase Modularity and Interpretability of Automated Essay Evaluation and Student Feedback

Chris Roche[1], Nathan Deinlein[1], Darryl Dawkins[1], Faizan Javed, PhD[1]
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
croche@smu.edu
ndeinlein@smu.edu
ddawkins@smu.edu
fjaved@smu.edu

Abstract. For English teachers and students who are dissatisfied with the one-size-fits-all approach of current Automated Essay Scoring (AES) systems, this research uses Natural Language Processing (NLP) techniques that provide a focus on configurability and interpretability. Unlike traditional AES models which are designed to provide an overall score based on pre-trained criteria, this tool allows teachers to tailor feedback based upon specific focus areas. The tool implements a user-interface that serves as a customizable rubric. Students' essays are inputted into the tool either by the student or by the teacher via the application's user-interface. Based on the rubric settings, the tool evaluates the essay and provides instant feedback. In addition to rubric-based feedback, the tool also implements a Multi-Armed Bandit recommender engine to suggest educational resources to the student that align with the rubric. Thus, reducing the amount of time teachers spend grading essay drafts and re-teaching. The tool developed and deployed as part of this research reduces the burden on teachers and provides instant, customizable feedback to students. Our minimum estimation for time savings to students and teachers is 117 hours per semester. The effectiveness of the feedback criteria for predicting if an essay was proficient or needs improvement was measured using recall. The recall for the model built for the persuasive essays was 0.96 and 0.86 for the source dependent essay model.

1 Introduction

Essay writing is an essential part of education that is taught to students of all ages. It is a critical skill that has been demonstrated to improve both understanding and retention of material by students, especially when combined with other skills such as reading or mathematics (Graham S. G., 2013). Everyday millions of essays are scored by hand by teachers. This is a time-consuming process. A survey of over 40,000 teachers, kindergarten through twelfth grade, conducted by Scholastic in 2012 reported that teachers spend on average, three hours (Scholastic, 2012) and seventeen minutes working outside of standard school hours each day. Much of that time is spent by teachers either at home or in libraries grading by hand (Scholastic, 2012).

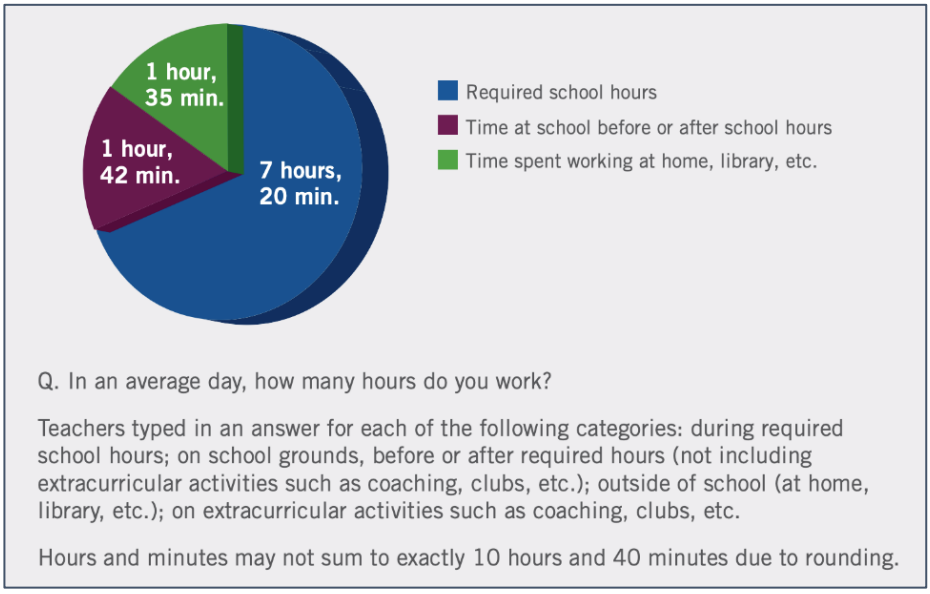

Figure 1: Teacher Workdays, (Scholastic, 2012)

While it is imperative writing serves as a foundational skill for students in kindergarten through twelfth grade, as many as one in two teachers (47%) nationally have expressed that they feel ill-prepared to teach and evaluate writing (Kiuhara, 2009). In the same survey by Kiuhara, et. al., the majority of teachers stated they made special adaptations for struggling students infrequently, where infrequently was defined as less than once or twice per month.

Students who struggle with writing skills are at a disadvantage when it comes to mastering the subject and demonstrating what they learned (Radloff 2001). Whereas students who do have strong writing skills can articulate what they learned or articulate their ideas through writing. Allowing them to actively engage by note-taking, summarizing, and journaling puts them at an advantage when it comes to mastering the subject (Radloff 2001). Not only do poor writing skills make it difficult to succeed in the classroom, but it also makes landing a job more difficult. Effective writing skills are deemed a necessity by employers for success in the workplace, and some have missed a job opportunity due to inadequate writing skills (Radloff 2001). To ensure a student's ability to contribute to society to their fullest potential, the tool aims to help those who are struggling, so they too can master the subject and avoid missing any employment opportunities.

Increasingly, schools are turning towards Automated Essay Scoring (AES) systems such as eRater, Intellimetric, and Grammarly to assist teachers. These systems provide a top-level score to students and rudimentary feedback in areas such as spelling and grammar. To date, most of the research into Automated Essay Scoring has been solely

on improving the accuracy of the models and not on feedback to students. Studies show that the least useful form of feedback for students is spelling and grammar ((McNamara D. C., 2013). The most effective feedback is in more complex areas, such as thesis sentences, the contextual flow of the essay, and drafting.

The research has two goals: to reduce the burden placed on teachers to grade multiple essays, while simultaneously improving student education. To achieve the first, the tool will provide the foundation for a modular NLP system that empowers teachers to customize the criteria based on the assignment. The tool will provide a set of pre-configured grading criteria that can be included or excluded from consideration during feedback. If the teacher feels that the tool is too lenient or too strict for a particular criterion, they can adjust the threshold accordingly for what prompts feedback. Both the teacher and the student can benefit from the feedback and metrics. Reducing the time to grade and improving the quality of education for each, respectively.

For instructors it will reduce time to grade the assignment by providing the teachers analysis of the essay prior to grading it, so they can focus on evaluating what the model can't. For students, it will improve the quality of their education by instantly and continuously providing actionable feedback based on expectations for that assignment, as opposed to having to wait for the teacher. A student will be able to submit their essay to the automated evaluation model as many times as they like prior to submitting their final draft to the teacher.

Secondly, this model aims to improve student feedback and growth. Using a recommender engine, the system will provide recommendations to external educational resources where the student can learn to improve upon the specific areas of concern in their essay. Students will be prompted to provide feedback on the relevance and quality of the resources and the recommender engine will adapt over time. This will automate the process of providing students with re-teaching opportunities which could normally only be provided by the teacher. The recommender engine will leverage the feedback given to the student from the NLP application to suggest video content to the student. This will allow for automated reteaching of problem areas for the student. Use of this part of the application should reduce not only time spent after school tutoring and re-teaching, but also time spent in the classroom providing instruction for the individual needs of students.

The goal of this model is to serve both teachers and students by supplementing the grading process for written assignments, it is not to implement an end-to-end solution. As an extensible framework to the grading process the model will speed up the grading process for teachers, provide students with instant feedback on their essays, and take advantage of re-teaching opportunities that would have been missed otherwise.

2 Literature Review

The literature review focuses on three primary areas: a survey of existing uses of Machine Learning and NLP to assist teachers and students, their shortcomings or perceived gaps, and emerging solutions.

2.1 Existing Student Aid Tools

While automated grading systems have been helpful to teachers, other tools are currently being researched and utilized to further aid students in improving their abilities. Grading a written response can be very subjective; to minimize the subjectivity involved in grading an essay, when teachers assign an essay to students, they also provide them with a rubric.

Rubrics are designed by the instructor for the student, and they minimize subjectivity by outlining the expectations for the assignment. The rubric will commonly contain the expected sections, what criteria the teacher will use to grade, and what percentage of the overall grade those criteria are worth. As the student works on the assignment, they know what the teacher is expecting and how the essay will be graded. A rubric also provides a means for communicating feedback to the student by identifying areas, in terms of the set criteria, where the student has excelled or underperformed.

The ability to highlight areas where the student has excelled or underperformed is critical to communicating what the student is doing correctly and where they need to improve. Knowing where a student needs to improve helps a teacher to properly instruct the student on how to improve their writing skills.

Currently, there is a major push to move from single score calculation towards either:

1. Rubric based feedback
2. Natural Language Generation (NLG) feedback based on specific user deficiencies in their writing.

This is believed to be more helpful to the student than single scoring options or models that only focus on grammatical errors (Graham S. &., 2007) (Woods, 2017).

The next progression in this area of research has been to combine this into a virtual tutor application (Mathew, Rohini, & Paulose, 2021) (McNamara D. C., 2013). These virtual tutor applications have utilized in training games and Response to Text Algorithms, which will be discussed in the next section, to attempt to improve students support in their writing or utilize NLP to function as a web scraper to find and answer questions asked of it (Mathew, Rohini, & Paulose, 2021) ((McNamara D. C., 2013).

Among the more interesting web scraping applications reviewed was an application based around improving student abilities in computer science. The application web-scraped college courses and websites to feed suggestions to the student to help advance

their computer literacy skills (Vo, et al., 2022). Though this does not directly correlate with improved grade school writing skills, it does open an opportunity to synthesize this framework with existing NLP practices for improved writing to create an application that not only reviews student work but suggests how to improve it. Also, this approach differs from most other existing self-help applications for writing, which simply search for an answer to a prompted question. Utilization of (Clopper CG, 2006) his recommendation for further learning technique, for English literacy and K through 12, specifically, was not found during the discovery part of this research.

2.2 Natural Language Processing and Education

NLP has recently opened many new avenues to assist students writing ability. However, NLP has only recently begun being deployed in the education field. To date, use cases have been limited and only moderately effective in assisting teachers with increasing the writing abilities of their students and in cutting down the time needed to review the essays by a teacher.

Early uses of NLP in the education sector focused on limited areas such as grammatical correctness and spelling issues (Litman, 2016). Many are also aware of its use in anti-plagiarism with applications such as turnitin.com. Automated essay scoring was looked at as the cornerstone of this technology allowing teachers to spend less time grading essays. Algorithms are used in this case to look at errors in parts of speech used, as well as sentence structural issues, for example, sentence fragments or run-ons (McNamara D. C., 2013). From here more complex NLP techniques have been applied to try to do more than simply look at these basic areas of language analysis.

Looking at lexical diversity was useful in classifying essays from students by linguistic level and sophistication (McNamara D. C., 2013). This initial look into linguistic sophistication was later used to build in features like sentence similarity coefficients to determine if an essay flowed correctly and stayed on topic (Zupanc, 2017). These similarity networks looked at items like thesis sentences and used lexical diversity to determine if appropriate support was given to the essay's main topic. Currently, lexical diversity is the highest level of analysis that is being done. Others have attempted to answer this question using varied techniques. The most common technique being semantic and thematic recognition algorithms (McNamara D. C., 2013) (Oyebode, et al., 2021) (Rahimi, 2017). The other alternative used frequently is a Response to Text Algorithm (RTA) which looks at similarities between a source given by the instructor and the students' work (Patout, 2019) (Zhang, 2019).

The above techniques show the initial progress of NLP into the education sector, however, there are several limitations to the current applications. Depending on the algorithms used, inefficiencies in the code or technology behind them can be limiting. The use of new Spark-style frameworks for NLP and new Spark NLP toolkits can help to bridge this issue ((Thomas, 2020).

The second issue is that most of the original tools created to assist teachers focus only on grammar and spelling or are based on essays created by professionals instead of students (Litman, 2016) (McNamara D. S., 2015). This means that most techniques look at the two least helpful sections of a student's essay in terms of making them a better writer (Litman, 2016), and algorithms designed to assist with more complex functions often have training sets that do not reflect actual student writing. The areas most in need of personalized attention are evidential support for the thesis, sentence flow, and semantic correctness of the essay (Litman, 2016) (McNamara D. S., 2015). Unfortunately, this is the most difficult area to create NLP frameworks for. Tools like the Response to Text Algorithm (RTA) are limited to the question asked of a given text and therefore have very limited use in a research paper writing setting (Rahimi, 2017) (Zhang, 2019).

There is also a lack of flexibility in the programs surveyed. Most programs can perform a single task, e.g., using semantic or thematic analysis to look at a single deficiency in the writing (Zupanc, 2017). Only a few applications have attempted to employ a modular framework to allow the application the flexibility to help teachers and students in more than a single area (Burstein, 2000) (Woods, 2017). However, the gap in the need for reteaching or personal tutoring has been left largely unaddressed by this modular framework as well. It is just starting to be met with varying degrees of success in a few applications that provide NLG feedback but no real reteaching (Litman, 2016) (Mathew, Rohini, & Paulose, 2021) (McNamara D. C., 2013).

Pointing out these gaps, this research seeks to balance these deficiencies in approach by focusing on three major areas:

1. Application flexibility for ease of use by teachers and students
2. Useful tutoring and reteaching through a recommender engine
3. The modularity of NLP processes to look at multiple deficiencies in student writing discovered in actual student essays.

3 Methods

3.1 ClassMate System

ClassMate, pictured in Figure 2, is comprised of several NLP models and a Recommender engine. Users can interact with the system via the front-end user-interface. The NLP engine receives the rubric criteria, student essays are input, then it provides feedback as output. The Recommender engine suggests educational videos based on the feedback provided by the NLP engine.

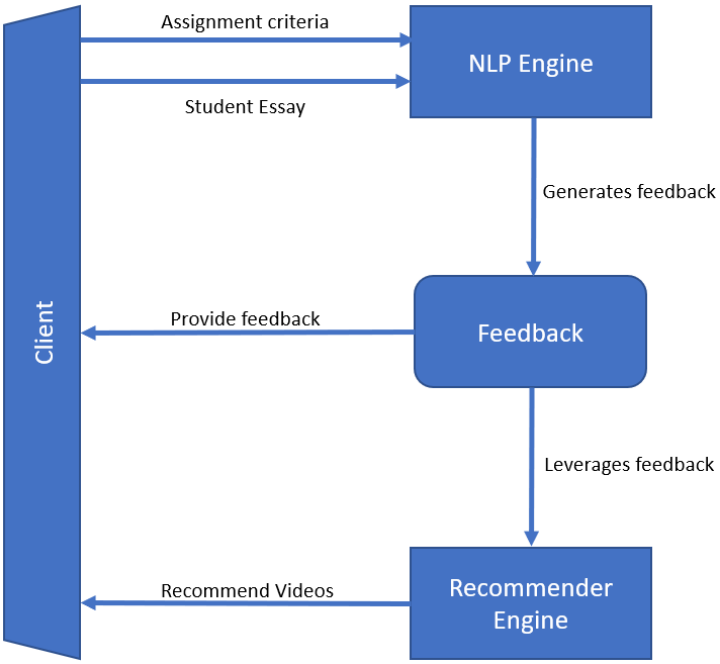

Figure 2: Architecture of the ClassMate system

3.2 Data

The data set contains over 17,200 essays. Each essay is between 150 and 500 words, written by a student from grades seven to ten. The essays are each scored by two independent teachers. The data is contained in both comma-separated value (CSV) and tab-separated value (TSV) files and available via the Kaggle Application Programming Interface (API) with the following command:

```
kaggle competitions download -c asap-aes
```

Tables 1 and 2, below, contain excerpts from randomly selected essays contained in the dataset. For brevity, they are not contained here in their entirety. Note the data set obfuscates personally identifiable information with tags beginning with the @ symbol. The information was removed by Kaggle using the Stanford NLP Group's Named Entity Recognizer (NER).

Table 1: Sample Essay, One

"Dear reader, @ORGANIZATION1 has had a dramatic effect on human life. It has changed the way we do almost everything today. The most well know, is the computer. This device has allowed people do buy things online, talk to people online, and also provides entertainment for some people. All good qualities that make everyone's lives easier. Imagine you look into your refrigerator and you notice it's almost empty. Someone is using the car and you need to go grocery shopping and the store is too far. What do you do? Well you could go on a computer and look for food online..."

Table 2: Sample Essay, Two

"In the excerpt "The Mooring Mast" by author @ORGANIZATION2, the builders of the Empire State Building faced some obstacles in attempting to allow dirigibles to dock there. A first obstacle the builders faced was the concern of the building collapein over time with "all the weight of the dirigibles. They would have to spend more money and time to create a frame for the Empire State Building to support the dirigibles." @CAPS1 @NUM1 supports that obstacle "a thousand foot dirigible moored at the top of the building would add stress to the buildings frame". A second obstacle builders of the Empire State Building came upon on attempting to allow dirigibles to dock at the building was the concerns of the thousands of citizen just below the tall building..."

Note: all spelling and grammatical errors in the above excerpts are as they exist in the data set.

Of the 8 essay corpuses included in the Kaggle set, 5 were written by 10^{th} graders. The 5 essay corpuses written by 10^{th} graders were used to train and test the models. From the 10^{th} grade essays there were 2 persuasive corpuses, with different prompts, and 3 source dependent essay corpus with different prompts. Scores from the essay ranged from 1-4, 1-6, and one essay corpus ranged from 0 - 60. An essay was considered a success if the student scored greater than 2, greater than 3, or greater than 40, respectively.

3.3 Experiment

The experiments performed were designed to provide evidence that the features used by the system rubric criteria are useful in deciding whether the quality of a student's essay is proficient or needs improvement. Student essays from the Automated Student Assessment Prize (ASAP) data set from Kaggle were used to perform the analysis. The data is from a 2012 competition hosted by the Hewlett Foundation.

3.4 Natural Language Processing

NLP was used to analyze student essays to create features for the prediction model and generate feedback to the user. The following sub-sections provide descriptions of the NLP features used.

3.4.1 Features & Feedback

Vocabulary. One rubric criterion implemented as part of this research was a measure of the essay's vocabulary diversity. If the essay submitted for feedback was found to be below the median diversity of the essay set for a given grade level, the tool presented vocabulary as a potential focus area for the student. We defined vocabulary diversity as the total number of words in the essay divided by the number of unique words in the essay. This was converted to a percentage.

When determining the median of the essay set, we weighted each training essay's diversity based on its provided score. I.e., the diversity of a training essay assigned a score of five was weighted five times as much as a training essay with a score of one when computing the median.

If the submitted essay was found to have a vocabulary diversity below the median for the given grade level, the tool took three actions:

1. The student was informed of their vocabulary score
2. The student was informed of the two most frequently used words found in the essay and prompted to consider using a thesaurus
3. The student was presented with a link to a learning resource on thesauruses

Figures 3 and 4 below show example reports to the student for the vocabulary criteria:

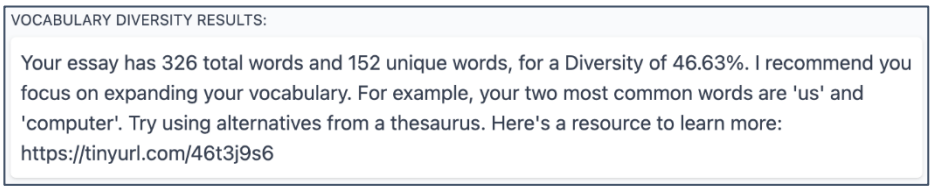

Figure 3: Example Vocabulary Diversity, below the median

Figure 4: Example Vocabulary Diversity, above the median

Word Count. Number of words per document is a very simple statistic to collect, yet a foundational metric in evaluating a student's essay. It can be presented to the student on its own or used as the basis for generating more complex feedback like vocabulary diversity. The rubric allows the teacher or student to set a minimum word count requirement so that the students will continuously be made aware of their progress towards meeting the requirement.

Sentence Count and Average Sentence Length. Sentence count per document is another part of the foundational metrics collected by the research. This feature alone can provide useful insight to the teachers prior to grading the essay and give them a sense of what to expect when grading the writing assignment. Combined with word count, appropriate feedback can be provided. For example, if a student met the word count requirement but sentence count is lower than expected, the tool would provide feedback to the student such as, "Your word count looks good, but due to the low number of sentences you may want to check your essay for run-on sentences."

Inversely if a student met the word count requirement but sentence count is higher than expected feedback to the student would be along the lines of, "Your word count looks good, but due to the high number of sentences you may want to check your essay for fragmented sentences." The threshold for a "high" number of sentences versus a "low" number of sentences is pre-defined by the tool based on data derived from the data set.

Average sentence length per document will help the tool re-enforce whether the essay needs to be checked for run-on or fragmented sentences. Especially in the case where word count and sentence count alone do not indicate a run-on sentence or fragmented sentence.

Extractive Summarization. The extractive summarization criterion relies on the teacher or student to input a prompt or key words for the assignment. The tool identifies and ranks the sentences from the essay that are most like the prompt or key words, using vector similarity. The rank serves as the quantifiable metric on whether the tool provides feedback to the student. The threshold for whether the tool provides feedback can be adjusted by the teacher in the event they feel it is too sensitive or not sensitive enough.

Once the tool identifies an essay that has a low similarity in relation to the prompt or key words, it will notify the student that they are not staying on topic. Additionally, since sentences are ranked, it quickly identifies which sentences are most in need of attention.

Recommender Engine. For the reteaching part of this application, we will be using videos from YouTube, Khan Academy, and similar resources to have ready-made content for the students to view. This will assist in retraining and boosting future scores

based on the deficiencies found in their essay. A list of videos will first be assembled and sorted into the categories of deficiencies. Then once a student submits an essay to be analyzed, the area of greatest deficiency will be put forward as the area to improve. This will then run a multi-armed bandit reinforced learning algorithm on the videos of that category.

The multi-armed bandit is a very basic reinforced learning algorithm. Its goal is to balance experimenting with different video options while also coming up with a suggestion for the most helpful video for each category. It does this by displaying a video and then receiving student feedback about said video. Based on the rating the student gives, a rank is assigned to the video. As more students watch the videos and give a rating, the algorithm "remembers" which videos received the highest rankings and will display those a set percentage of the time. There is also a feature to pick a random video. This allows for other videos to have a chance to be ranked and have a chance to become the top ranked video. We did this so that when new videos are added to the lists by teachers in the future, those videos would have an equal chance to then become the top ranked video.

There are three of these algorithms that power our recommendations, one for each category. If the NLP models determine a deficiency in the students writing, the UI will then display the URL suggested for that deficiency.

3.6 Model Evaluation Metrics

Recall was used to evaluate the machine learning models, since the cutoff for a success or fail was based on a below or above average score. It was decided that it is best for the model to incorrectly flag an essay as a failure as opposed to incorrectly flagging an essay as a success. The cost of providing a student with feedback and learning material is inexpensive compared to missing the opportunity to provide feedback to a student that needed it. For that reason, false positives were preferred over false negatives.

3.7 Modeling

Two logistic regression models were created, one for the source dependent essays and one for the persuasive essays. Essays with 3 or less sentences were not included in the machine learning experiment, but word count and sentence count analyses were still performed. Nine features were included in the machine learning models and displayed below in table 3.

Table 3: Features used for machine learning models

Features

Word count	Median Key Rank
Sentence count	Mode Key Rank
Average Sentence length	Average Key Word Count
Lexical Diversity MLTD	Mean Key Lexical Diversity
Average Key Rank	

A ten percent holdout set was created using a stratified split, yielding a training set with 6,812 records comprised of 19% success records. The holdout set had 757 records comprised of ~18% success records. All the training feature data was scaled and fitted, and the holdout set feature data was just fitted using the same scalar from the training data.

Fivefold cross validation was applied to all the training data. The metrics for each of the five folds were consistent, yielding a variance of 0.50 and 0.81 in recall for persuasive and source dependent models, respectively. Since the variability was low, the threshold for a model was tuned using the entire test set. The final model was selected based on precision and the model with the best precision had a threshold of 0.3.

4 Results

The results for the final persuasive logistic model yielded a recall of 0.96. The results from the source dependent model yielded a recall of 0.86. Word count, key rank median, lexical diversity, key mean lexical diversity, and average sentence length were the top 5 features for both models. Indicating that if a student increases their word count, improves their diversity and/or stays on topic their odds of passing will increase.

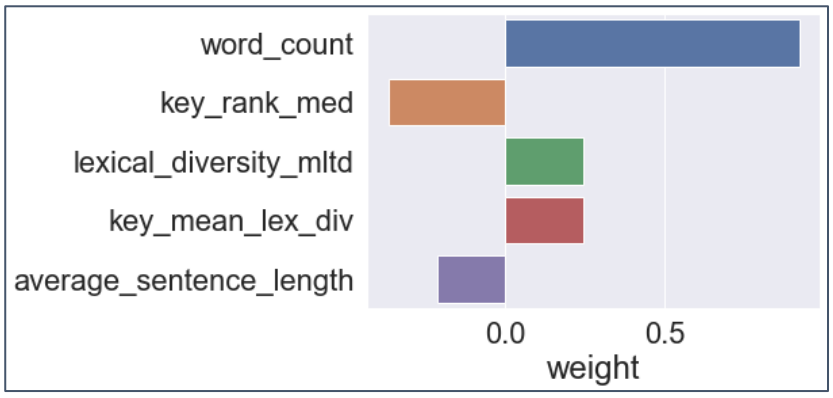

Figure 5: Feature importance for persuasive model

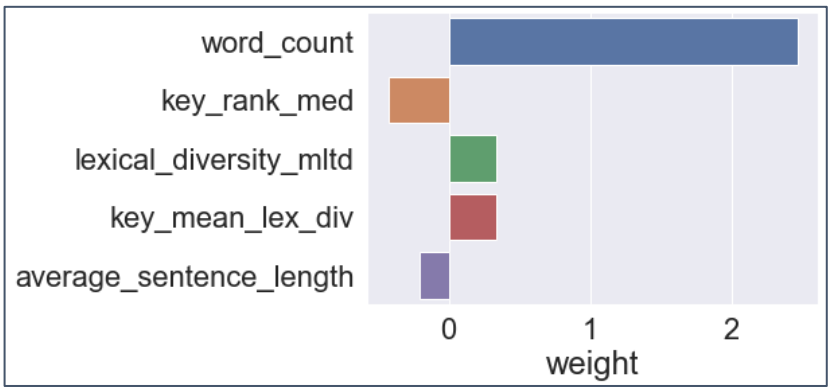

Figure 6: Feature importance for source dependent model

While no teacher or workload is the same, we can gather some baseline business statistics by using the data collected by Scholastic shown above in Figure 1. According to the data teachers spend an extra three hours and seventeen minutes a day working on items outside of their school hours. If only a third of that were spent on grading student essays, that would still be a little over an hour a night spent grading papers alone. The US average high school class size is 23.3 students (National Center for Education Statistics, 2019). Many high school teachers teach at least five classes a day allowing us to expect them to have around 100 to 125 students a day. If a teacher spent two minutes looking at each submission, that would mean a minimum of three hours to go through every paper. While this minimum is unrealistic to judge any teacher by, we will use it to explain the improvement to the teacher's and student's life that this application can bring. Depending on length of the student writing assignment, the turnaround for feedback to the student composition can vary, but they will have feedback in under a minute versus one to multiple days waiting to receive feedback from the instructor. This allows students to quickly begin reworking their assignment, cutting down on the timelines a teacher needs to plan for the traditional feedback loop to complete its cycle.

Students can also submit as many times as they want. If each student were to use the assessment twice in our theoretical class, the teacher would need a minimum of 6 hours to grade all the submitted papers. Instead, the instructor can use those 6 hours for other purposes, i.e., classroom prep, coaching, tutoring, or relaxing. This number rises if students proceed to turn in papers more than the two times discussed in this example.

This is not the only function of the application, however. The application also creates an individual relearning plan for the student based on the selected criteria, using a recommender engine. The average length of the videos provided by the recommender engine is around nine minutes. A student, currently, could receive up to three suggestions, creating a range of 0 to 24 minutes of additional instruction for a single pass on their paper. If we take a median time frame of 12 minutes. It would take a teacher three hours of grading and 20 hours to cover the reteaching material with each student individually. Taking this relatively conservative estimate and use five writing

assignments a semester as another low estimate, the average teacher would be saved 117 hours, or nearly five days, a semester. This is a huge amount of time saved for the instructor at this low end of the estimation. This number is likely to be much higher in actual practice depending on the teacher's actual usage.

Table 4: Chart of time savings (hours) based on 12-minute reteaching and 100 students per semester

Number of uses per paper	5 papers per semester	7 papers per semester	10 papers per semester	16 papers per semester (weekly)
2	233	327	467	747
3	350	490	700	1120
4	467	653	933	1493
5	583	817	1167	1867

5 Discussion

As discussed in the literature review, this research offers the first combination of dynamic teacher input, NLG student feedback, and use of a recommender engine for assisting students with improving their English composition skills. Current NLP models have mostly, up until now, given a feedback score with little explanation as to how that score was calculated. By giving the students constructive criticism, like the hand grading performed by a teacher through use of a rubric, this system allows students to improve their future writing. This was coupled with a dynamic rubric selection feature, allowing teachers or students to focus on a specific attribute or attributes of the writing process they wished to see improvement in.

5.1 Identification and Interpretability of Student Writing Deficiencies

The result of this research is a fully usable tool for students to input essays and receive feedback. The User Interface was created with the Gradio Application Programming Interface (API). When accessed from a browser, the user interface presents a text box for inputting the essay. Different rubric criteria are presented as optional check boxes and control which NLP models are executed against the input text. Additionally, the user can select their grade level. On the back end, grade level is used to determine the threshold for specific metrics. For example, the median vocabulary diversity of each grade level in the data set was computed. When the user selects tenth grade, their feedback will be evaluated based on how they compare to the metrics derived from other essays labeled as tenth grade.

The figure below shows a sample of the student input panel in action.

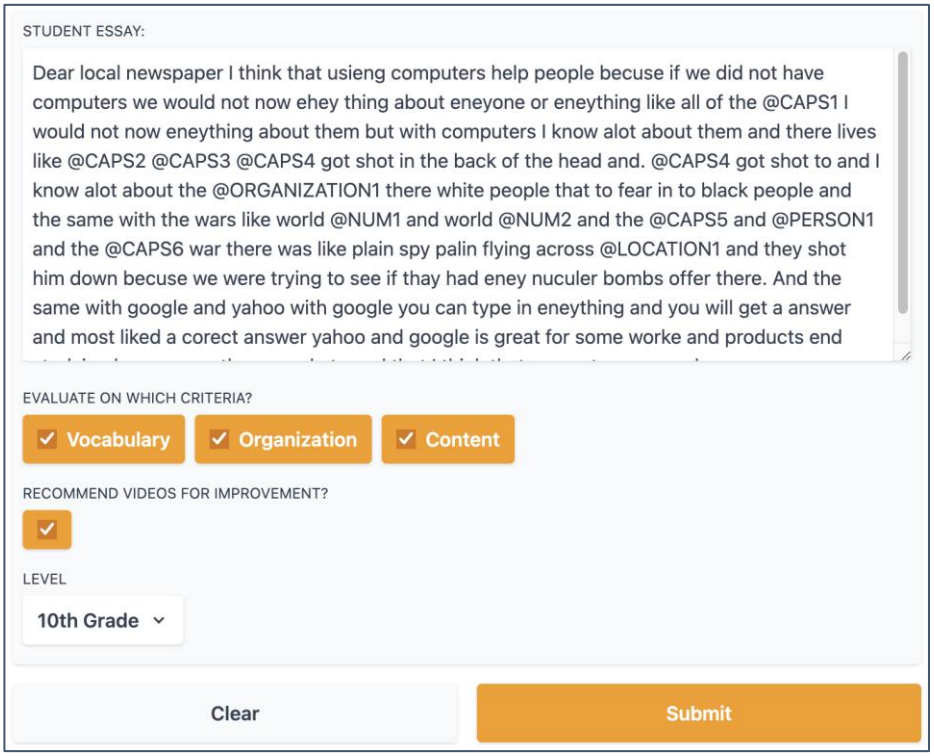

Figure 7: Essay input panel with sample rubric selections

Figure 8 (below) shows the user interface output from the essay shown in figure 7 (above). As can be seen, the tool determined the input essay required attention for the vocabulary diversity criteria but that the others were sufficient.

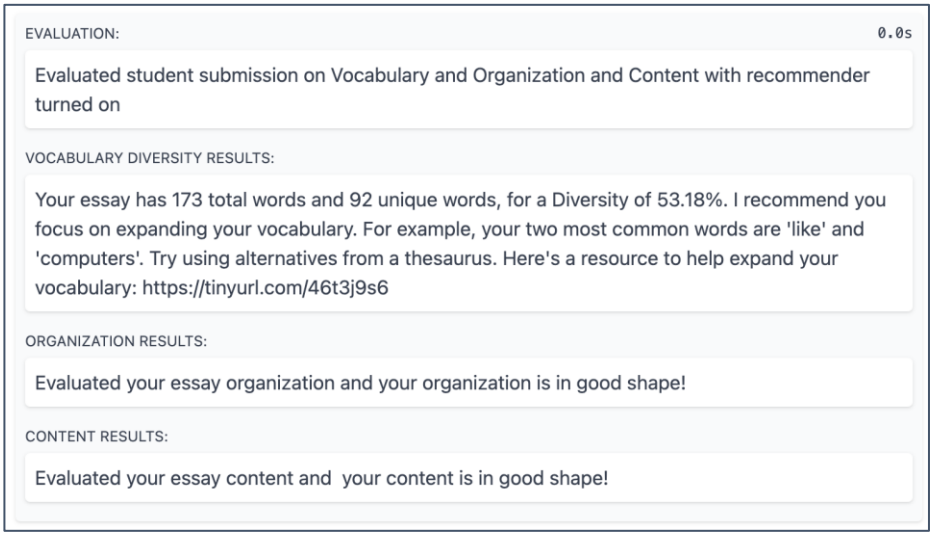

Figure 8: Essay feedback output panel with sample rubric selections

5.2 Recommender Engine for Student Feedback

When paired with the recommender engine, this is the first time that students can receive additional instruction in an area without intervention from the teacher. All these attributes go a long way in assisting with cutting down on teacher grading time, lag time in students waiting for feedback, and cutting down on reteaching time during or after class for the students and teacher.

The recommender engine was a large part of the undertaking of this research. The goal was to design a dynamic and useful tool for students that would take the workload off teachers to reteach certain materials in or out of class. There were two main features that were taken into consideration when designing the recommender engine: student feedback and continuous improvement.

With a recommender engine, the basic idea is to use feedback to display the best result as much as possible to help students. Giving students the ability to respond after each video with how helpful the video is, gives the system the ability to provide the most useful videos to students. If a video is found to not be helpful, it will be displayed fewer and fewer times. This also helps with the second feature, continuous improvement. The goal here was to do two different things, first report back videos that were not helpful, so that they could be removed from the lists of possible options. Second, to allow teachers to submit additional videos they use for reteaching these subjects. As these new videos get added in, the recommender engine would start to show these to students to receive feedback on their usefulness. As the new videos became more widespread in the exploration part of the recommender engine, their positions would be reevaluated

with the least helpful being dropped from the list of recommendations and the helpful ones being kept on. This cycle of pruning would always keep the engine cycling through a small group of new videos and the most helpful videos for each content area.

Figure 9 (below) shows an example of the Recommender Engine displaying a resource link to the student as part of the essay feedback output panel:

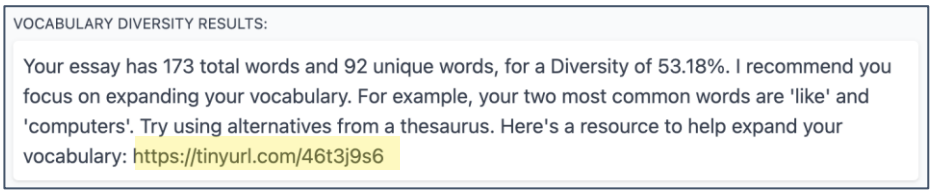

Figure 9: Recommender engine suggests a video for a student needing to expand vocabulary

Once the user clicks on the resource link, they are redirected to an embedded video player and presented with an educational video related to their specific improvement needs and a form to rate the usefulness of the video. As discussed above, this allows the tool to continuously improve recommendations.

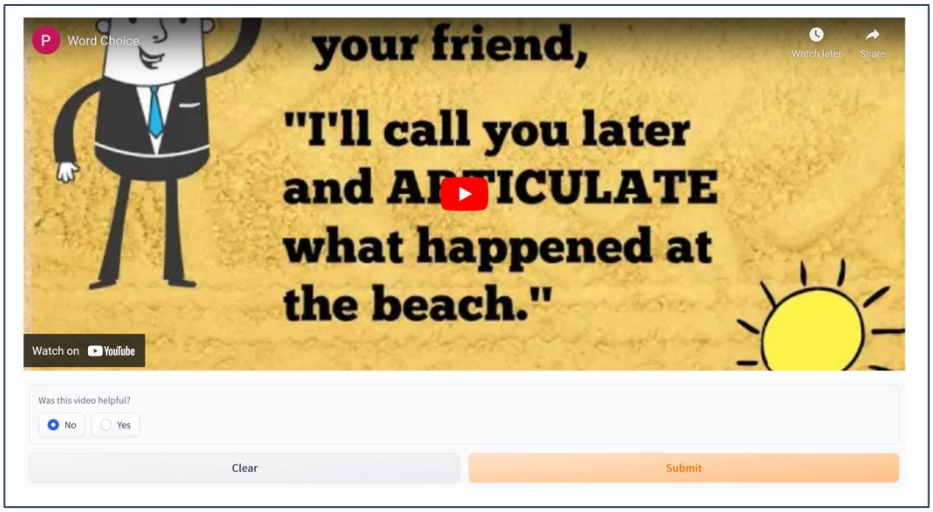

Figure 10: Recommender engine panel to receive feedback on how useful it was

5.3 Ideal Consumer and Typical Use Case

The ideal use case would be for teachers to implement and administer this application in their classroom to supplement the grading process. Teachers would be expected to

provide the rubric criteria and input the criteria into the model. The intent would be for both teachers and students to use and benefit from the application. The teachers would use it to expedite the grading process and the students would use it throughout the writing process to get instant feedback on their papers. This application also needs to be administered in a situation where students and teachers have a good technological literacy as well as English literacy. The application allows for a wide range of users and helps democratize literacy by being usable in all web browsers as well as mobile platforms. This would allow for the application to be utilized by schools with limited funding that typically have iPads or Android tablets instead of full laptops.

5.4 Ethical Considerations

The impact of any use of NLP on the educational environment should be scrutinized at the highest level before implementation. Because any tool used in the classroom can have both positive and negative impacts to the students who use it, care should be given to fairness and other ethical concerns when testing and implementing it in the classroom. The first hurdle is that the application itself will suffer from a bias native to its creators. English is an extremely complex language with 4 main dialects just in the United States alone with many subdialects from each main dialect group (Clopper CG, 2006). The consequence of this is that there is no standard English that is in practice and used by all regions in the United States. These dialects have differences in syntax and word usage that might cause any NLP tool to identify deficiencies in the writing differently. This is something that needs to be acknowledged and understood as implementation goes through. Schools that have English as a Second Language (ESL) students, will also tend to have similar issues as well. This does not mean that NLP should not be used in the educational setting, but that instructors that plan to use it as a tool should be aware that further follow-up with these minority groups will be required. It was for this reason that the NLP solution proposed by this research attempts to avoid giving students a single score metric on which the student will be graded. By providing a rubric that provides areas where the student can improve and by looking at deficiencies less susceptible to this sort of bias, the solution is able to circumvent some of the ethical issues discussed.

The second ethical issue is on implementation of the tool in a classroom setting. The use of any learning tool should be looked at continually and reassessed. This is because if the tool ends up being a hinderance or not as good as the alternative solution, serious damage can be done (Ferguson H, 2007) to a child's development. Many studies have been done on the detriments to a student due to a poor learning environment (Ferguson H, 2007). Poor academic achievement can lead to the student losing confidence, avoidance, and antisocial behaviors within the classroom as the feeling of disconnection grows (Ferguson H, 2007). These can be heightened by the fact that this solution would create an additional layer between the student and teacher, where the student is submitting their work to a program and not getting actual feedback from the teacher.

While many schools are testing hybrid learning styles and online learning alternatives to the traditional classroom (Annelies, 2021), the ethics of this should be considered when considering implementing an online tool as a solution.

Technological literacy should also be considered as a possible ethical issue. Around 16 percent of students, 11.3 million, live below the poverty line based on 2020 numbers (National Center for Education Statistics, 2022). According to the USDA, 7.9 percent of households had food insecurity for the year 2020 (Coleman-Jensen, Rabbitt, Gregory, & Singh, 2021). These households are unlikely to be able to spend their resources on Technological literacy. Lack of competency in this area or lack of access to technology would make any computer-based solution hard to implement and would unfairly hinder those who find themselves in this sort of situation. Administrators and teachers, who look at our application as a solution to use in the classroom, will also need to look at their implementation from an ethical standpoint. Is implementation of the tool helpful overall? Are there accommodations that can be made for those students who find themselves lacking computer literacy, either due to poverty or other issues? If no accommodation can be made, will the efficacy of the tool be compromised?

Finally, the last major ethical issue regards responsibility for externalities of use of the solution. If a student is pointed in a direction other than that intended by the instructor, i.e., takes advice from the application that causes them to get a lower grade, with whom does that responsibility reside? Should it be with the creators of the application for pointing students in the incorrect direction, or with the instructor who in theory should be monitoring the progress of their student's essays? Because the grading of essays is subjective, any sort of misalignment between the goals of the solution put forth by this research and that of the grading teacher could cause issues with students receiving lower grades than they would have due to the coaching of the application. It should be noted that this can be negated by using a wide range of teaching materials included for selection by the recommender engine, as well as the ability of the teacher to select which deficiencies they specifically wish to cover using the applications rubric building selection functionality.

5.5 Future Opportunities

A shortcoming of the vocabulary diversity score was found to be the abundance of spelling errors in the data set. Because the research determined the score based on the number of unique words in an essay, if the same word was used multiple times but was misspelled in one or more of its uses, that word was factored into the calculation as being unique. A future improvement opportunity would be to pipeline a spell check and correction feature prior to determining the number of unique words in an essay.

In addition, for vocabulary diversity, this tool returns the two most frequently used words in an input essay when the diversity is determined to be below the median and the tool then makes recommendations for resources on the use of a thesaurus. A future

improvement would be to use a model such as Word2Vec to return specific recommendations for replacement words, in addition to referencing a thesaurus.

Another improvement in the future to improve the usefulness of the tool would be a logging mechanism on a per-student basis. This could serve two purposes:

1. Provide a usage report to the teacher. This would allow the teacher to monitor student engagement, spot-check the performance of the tool, and follow-up with the student if intervention is deemed necessary
2. Allow the NLP models to incorporate a student's historical data to further improve recommendations

Incorporating a student's historical performance provides exciting opportunities within the Machine Learning space. The tool presently only has access to the specific essay being evaluated and the metrics gathered from the data set. Access to a student's previous submissions and feedback would provide the opportunity to use algorithms that could spot trends. This would allow it to reinforce specific areas of focus that the student frequently has issues with as well as tailor the thresholds used to determine whether a specific criterion is recommended for improvement.

6 Conclusion

This research made the case for the need to improve Automated Essay Scoring (AES) and automated essay feedback models. Specifically, the goal of this research was to redirect the focus of such models from strictly grading to providing a platform for teachers to specify the criteria they wish the tool to take into consideration and then for the tool to provide learning feedback to the student based on their performance. The goal of this research was not to implement a complete end-to-end solution with all possible grading criteria, it was to implement a prototype solution in an extensible framework and demonstrate that it is a viable solution to reducing teacher burden and improving student education.

Through the development and deployment of a tool teachers and students alike can use, this research demonstrated that existing Natural Language Processing techniques are much more capable than simply assigning a score to an essay. Three different prototype criteria were implemented in the tool: vocabulary analysis, essay structure, and essay content. This allows students to submit early drafts of essays to the tool for initial feedback and select which areas they or their teachers want them to focus on. Thus, reducing the burden on already overburdened teachers. In addition to providing feedback to the students, the tool has a mechanism for users to provide feedback to it. Using a Multi-Armed Bandit Recommender Engine, the quality and utility of the feedback given to the students continuously improves.

Acknowledgments. The authors would like to thank our research advisor, Dr. Faizan Javed. We would also like to extend our thanks to our SMU faculty advisors, Dr. Jacquelyn Cheun and Dr. Tim Musgrove. Additionally, the authors wish to thank the many teachers who voluntarily evaluated the training data and helped determine which deficiencies to target, as well as providing educational resources to feed into the Recommender Engine.

References

Annelies, R. (2021). Exploring Student and Teacher Experiences in Hybrid Learning Environments: Does Presence Matter? *Postdigital Science and Education*, 1-22.

Burstein, J. &. (2000). Benefits of modularity in an automated essay scoring system. *Proceedings of the COLING-2000 Workshop on Using Toolsets and Architectures to Build NLP Systems*, (pp. 44–50).

Clopper CG, L. S. (2006). Perceptual similarity of regional dialects of American English. *J Acoust Soc Am*, 119(1):566-574.

Coleman-Jensen, A., Rabbitt, M. P., Gregory, C. A., & Singh, A. (2021). *Household Food Security in the.* U.S. Department of Agriculture.

Crossley, S. A. (2016). Incorporating Learning Characteristics into Automatic Essay Scoring Models: what Individual Differences and Linguistic Features Tell Us about Writing Quality. *Journal of Education Data Mining*, p1-19.

Ferguson H, B. S. (2007). The impact of poverty on educational outcomes for children. *Paediatr Child Health*, 12(8):701-706.

Graham, S. &. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, 445–476.

Graham, S. G. (2013). Writing: importance, development, and instruction. *Read Writ 26*, 1-15.

Kastrati, Z., Dalipi, F., Shariq Imran, A., Pireva Nuci, K., & Ahmad Wani, M. (2021). Sentiment Analysis of Students' Feedback with NLP and Deep Learning: A Systematic Mapping Study. *Applied Sciences*, 3986–.

Kelly, A. &. (2020). College in the Time of Coronavirus: Challenges Facing American Higher Education. *American Enterprise Institure*.

Kiuhara, S. G. (2009). Teaching writing to high school students: A national survery. *Journal of Educational Psycology, 101*, 136-160.

Litman, D. (2016). Natural Language Processing for Enhancing Teaching and Learning. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16).*

Mathew, A. N., R. V., & Paulose, J. (2021). NLP-based personal learning assistant for school education. *International Journal of Electrical and Computer Engineering*, 4522–.

McNamara, D. C. (2013). Natural language processing in an intelligent writing strategy. *Behav Res*, 499-515.

McNamara, D. S. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 35–59.

National Center for Education Statistics. (2019). *Average class size in public schools, by class type and state: 2017–18.* U.S. Department of Educatio.

National Center for Education Statistics. (2022, May 1). *Characteristics of Children's Families*. Retrieved from National Center for Education Statistics: https://nces.ed.gov/programs/coe/indicator/cce

Oyebode, O., Ndulue, C., Adib, A., Mulchandani, D., Suruliraj, B., Orji, F. A., . . . Orji, R. (2021). Health, Psychosocial, and Social Issues Emanating From the COVID-19 Pandemic Based on Social Media Comments: Text Mining and Thematic Analysis Approach. *JMIR medical informatics*, e22734-e22734.

Patout, P.-A. &. (2019). Towards context-aware automated writing evaluation systems. *Proceedings of the 1st ACM SIGSOFT International Workshop on Education through Advanced Software Engineering and Artificial Intelligence*, (pp. 17–20).

Rahimi, Z. L. (2017). Assessing Students' Use of Evidence and Organization in Response-to-Text Writing:Using Natural Language Processing for Rubric-Based Automated Scoring. *INT J Artif Intell Educ*, 694-728.

Scholastic, I. (2012). *Primary Sources: America's Teachers on the Teaching Profession.* Retrieved from https://www.scholastic.com/primarysources/pdfs/Gates2012_full.pdf

Thomas, A. (2020). *Natural Language Processing with Spark NLP.* O'Reilly Media, Inc.

Vo, N. N., Vu, Q. T., Vu, N. H., Vu, T. A., Mach, B. D., & Xu, G. (2022). Domain-specific NLP system to support learning path and curriculum. *Computers and Education: Artificial Intelligence*, 100042–.

Woods, B. A. (2017). Formative Essay Feedback Using Predictive Scoring Models. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 2071–2080).

Zhang, H. M. (2019). eRevise: Using Natural Language Processing to Provide Formative Feedback on Text Evidence Usage in Student Writing. *Proceedings of the AAAI Conference on Artificial Intelligence*, (pp. 9619-9625).

Zupanc, K. S. (2017). Evaluating Coherance of Essays Using Sentence-Similarity Networks. *Proceedings of the 18th International Conference on Computer Systems and Technologies*, (pp. 65-72).