The Perfect Essay Autograder

By: Andrew Kerr and Nathan Hill

Introduction

Teachers read hundreds of essays a year, repeating the same prompts year after year, which gets real boring, real fast. Furthermore, shifting though and staying consistent when grading each essay takes up time and effort that could be spent elsewhere. This problem has been solved for multiple choice, and even short response, questions already - through the use of an autograder. So, why not do the same for essays?

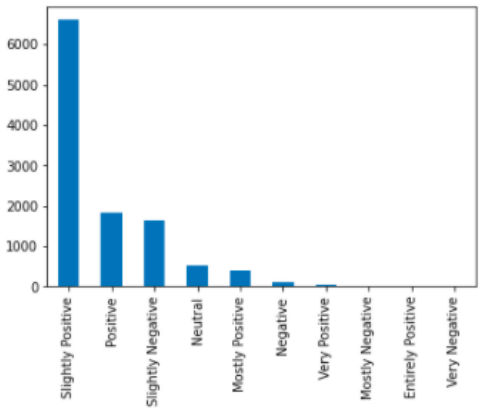
Data Collection and Cleaning

From the Hewlett Foundation on Kaggle, we acquired a dataset of over 11,000 essays written by students in grades 7 to 10. There were 8 different prompts, and these prompts can be classified as either expository or source dependent. Each individual essay has scores from two separate graders, however each prompt was graded on different point scales, making percentage the best way to compare scores rather than points. A huge part of writing is the ability to express emotion through words, and we were interested in whether essays with overall positive or negative tones achieved better scores and if the tone of an essay could help predict score. Therefore, we decided to pair these essay scores with polarity results from SENTIM-API. This API takes in a string, goes sentence by sentence, determines each sentence's polarity, on a scale from -1 to 1, and the polarity of the whole string, which turns out to be the average polarity of the sentences.

The provided essay data frame contained all of the prompts combined. Since we wanted to compare autograders for each specific essay to an autograder for all essays at once, we began by separating the essays by prompt. Upon considering the data, we decided not to include essays from prompt 2 since the graders used a different process for scoring, involving giving scores for specific aspects of the paper, and no grade overall. Next, we ran each essay set through SENTIM-API and merged the results with their corresponding essay and score. Finally, we merged all of these data frames together to make an overarching data frame, excluding essays from prompt 2.

Figure 1: Here the polarities are grouped in increments of 0.2, with 0 to 0.2 being slightly positive, 0.2 to 0.4 being positive, 0.4 to 0.6 being mostly positive, 0.6 to 0.8 being very positive, and 0.8 to 1 being entirely positive. The same labels are used for the negatives, and neutral values are those whose polarities are exactly 0.

```
Slightly Positive    6601
Positive             1832
Slightly Negative    1645
Neutral               518
Mostly Positive       385
Negative              109
Very Positive          59
Mostly Negative        17
Entirely Positive       6
Very Negative           4
Name: result.level, dtype: int64
```



Data Exploration

The first thing we wanted to know was the distribution of polarity scores essays received, and if they appeared to be consistent across all prompts. We quickly found that the majority of essays tended to be positive, but most of them were still close to having a neutral tone. We also discovered that each of the

prompts distribution of polarity scores slightly differed. For example, essays from prompt 4 tended to be mostly neutral, while those from prompt 5 were overwhelmingly positive, which could be due to the specific prompt of the essay. In response to these prompt to prompt differences, we decided to focus on the autograder for each prompt individually.

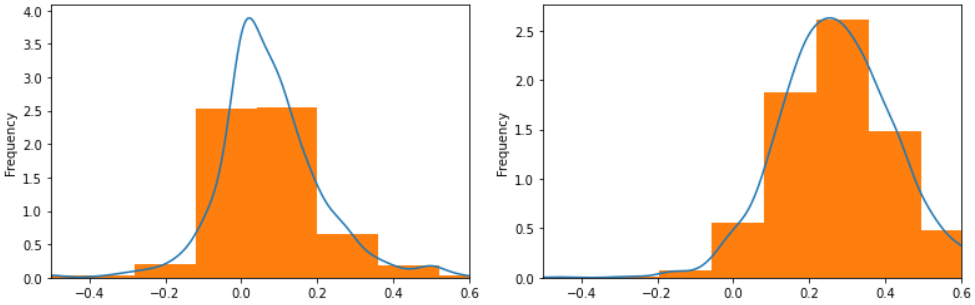
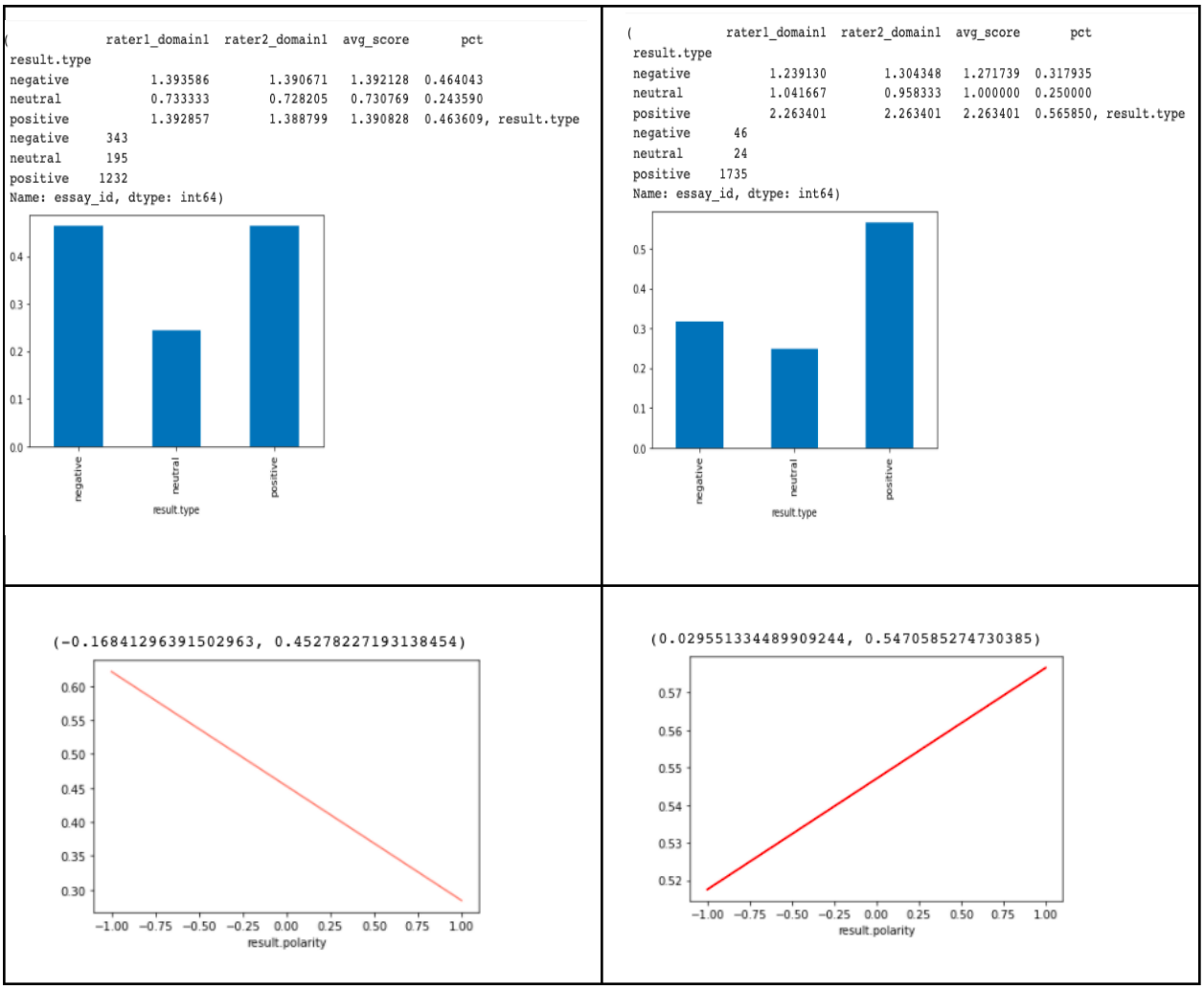


Figure 2: Histogram of polarity values for prompt 4 (left) and prompt 5 (right).

The variation in prompt to prompt polarity led us to investigate what affected the polarity of an essay. With the data we had, we were able to look at the grade level of the writer and type of essay. Essays from groups 1, 5, and 7 were written by 7th and 8th grade students, while those in groups 3, 4, 6, and 8 were written by 10th grade students. After separating the data for each group, we noticed that the middle school essays had an average polarity score of 0.1502, compared to an average of 0.0797 for the 10th graders. This was surprising since the average score was nearly double for middle school students. Next, we divided the essays based on prompt category; putting expository essays 1, 7, and 8 against source-based essays 3, 4, 5, and 6. Again, a difference emerged, as the expository essays had an average polarity score of 0.0857, meanwhile the source-based essays had an average score of 0.1275.

With this new information in mind, we shifted gears, investigating whether polarity scores are correlated with grades. We hypothesized that higher polarity scores might be associated with higher grades, due to the overall positive tone of the paper. To investigate this, we did two calculations for each essay group. First, we calculated the average score for each of the essays, grouped by whether they were negative, neutral, or positive. Next, we fit a linear regression between the polarity score and the score percent (score received / maximum possible score). To illustrate these data summaries we will compare and contrast essays in group 4, which was a source-based essay prompt written by 10th grade students, and essay group 5, also a source-based essay but written by 8th grade students.

Essay 4	Essay 5
The scores from essay prompt 4 strongly suggest that essays with neutral polarity scored much lower than both positive and negative, since the average neutral test score was 20 percentage points less than the average scores for positive or negative essays. In this instance, the data does not appear to be linear, so the regression line is not very useful.	Essay prompt 5 scores favored those who wrote more positive essays, and the vast majority of the students did. The regression line has a positive slope, which makes sense given the summary statistics.



Overall, essays that tended to get the highest scores were those that were either slightly positive or slightly negative, but not completely neutral or extreme on one side or another; not at all what we expected. It is impossible to say based on this data whether the graders preferred essays that were not neutral, or if students who are better writers tend to have slightly positive or negative polarity scores. Essentially, we cannot determine causation, but there is certainly evidence of a relationship between the polarity score of an essay and the grade assigned to it.

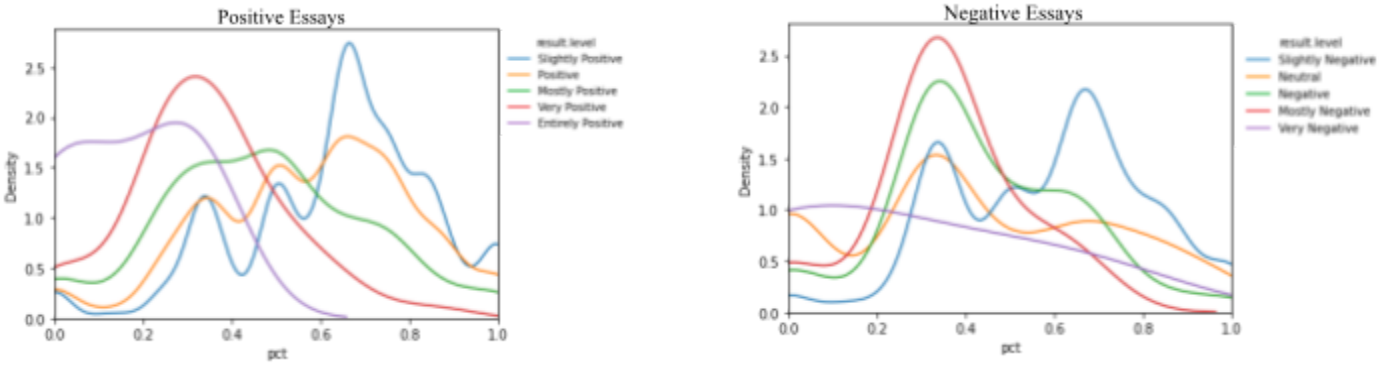


Figure 3: Density plots for each of the levels of polarity, divided in the same way as Figure 1. It is clear that the slightly positive and slightly negative essays received higher scores than those that were very positive or very negative. The plot is divided into 2 separate plots to make it easier for the reader to view each line.

Machine Learning

Based on our data exploration, the polarity score of an essay may not be the best method for predicting score. Thus, we created two K-Nearest Neighbors models: one predicting on the words in an essay and one predicting on the polarity score of the essay. For both models we ran each essay prompt individually, predicting the score, and all essays at once, predicting the score percentage. This is so we could compare a model grading on prompt specific essays to a general use model.

For the first model, each essay prompt was predicted on its corresponding term frequencies, weighted by the inverse document frequency (TF-IDF). We chose TF-IDF over purely TF because we did not want common words to take over the analysis. Conversely, when using all the essays at once we chose to use TF since we didn't want prompt specific words to be accounted for. When selecting the amount of neighbors to look at, k, we arbitrarily chose 10 since selecting based on the smallest test mean squared error (MSE) was not helpful.

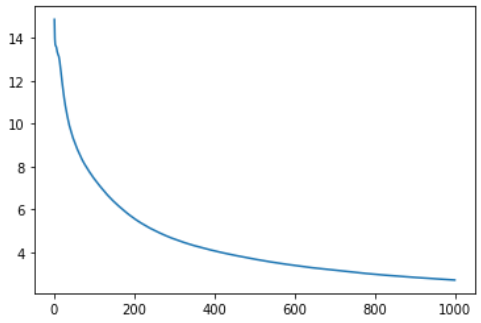


Figure 4:
K value for K-Nearest Neighbors against test MSE for prompt 1 essays

For the second model we used polarity score and polarity type (negative, neutral, positive) to predict the score and score percentage respectively. This time we were able to utilize Grid Search to find the best k value, resulting in k values of 87, 70, 82, 44, 131, 104, and 88 for each prompt and 219 for all together. Since we had trouble using MSE for the first model, but not this one, we concluded that the large amount of unique word combinations interfered with this algorithm.

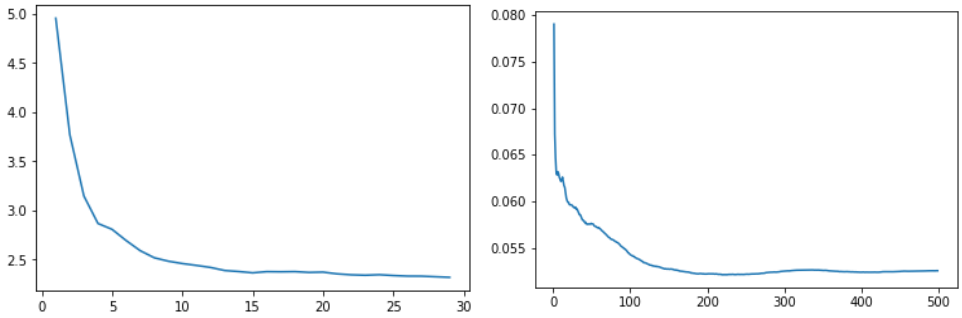


Figure 5:
K value for K-Nearest Neighbors against test MSE for prompt 1 essays (left) and for all essays (right)

When comparing the two models, on average the predictions from the polarity models performed better than those from the TF/TF-IDF models. This supports our findings from our data exploration; that the polarity does affect the score. However, in contrast to our expectations the general essay grader resulted in the smallest difference on average in predicted and actual. This is most likely due to the larger sample size of this data set as it contained all essays, while the prompt specific data sets contained much smaller sample sizes. Therefore, for our final model we decided to create a general grader, trained on all of the essay prompts, which computes the score percentage.

	TF/TF-IDF	Polarity
Prompt 1	2.5738	2.1558
Prompt 3	1.2903	0.863
Prompt 4	0.9704	0.6748
Prompt 5	2.0135	1.1207
Prompt 6	2.0713	1.3218
Prompt 7	4.0572	3.6454
Prompt 8	10.1981	9.7311
All Prompts	0.2688	0.2475

Table 1: Average difference in predicted score and actual score (predicted - actual) for specific essay prompt and average difference in predicted score percentage and actual score percentage for all prompts

Now that we know what we want to predict, we need to decide what feature to include in the model. We know we want to use polarity score and polarity type, but should we also include the grade of writer and type of essay? According to our data exploration, yes, however more is not always better. After looking at the MSE for each model combination, we decided to use all four features in our model since this model produced the smallest MSE at 0.0496. Thus, our final model, with a k of 40, resulted in a difference in average predicted score and average actual score of 0.1819, the smallest difference out of every previous model.

Conclusion

It turns out creating an all purpose essay autograder is difficult due to the enormous amount of factors affecting each essay. From the data we acquired, accounting for the grade of the writer, type of essay, the polarity, and the general polarity of the writing produces the best results. Furthermore, with the amount of samples we had, looking at all the essays overall provides better predictions than each prompt separately. With larger sample sizes of each prompt, we would expect the opposite conclusion since focusing on essays only from a single prompt results in less unexplained variability between the essays. Furthermore, we discovered that the polarity of an essay is correlated with different scores, as it proved to be a fairly good predictor on its own when compared to the TF-IDF model. The exact reason for this association remains unknown, but a possible explanation is that students who are more proficient at writing need to use less emotionally charged words to convey their message sufficiently, thus leading higher scores to be associated with moderate polarity scores. A follow-up experiment could be employed in order to specifically test this hypothesis, but it is impossible to determine causation from this data alone. Further research would allow teachers and students alike to take advantage of this, assisting teachers in the grading process and helping students understand what constitutes a well-written essay.