Assessing Student Support Effectiveness

A Major Qualifying Project

Submitted to the Faculty of

Worcester Polytechnic Institute

By: Benjamin Babalola

And Morgan Lee

Date: 3/25/2022

Project Advisor:

Professor Neil Heffernan

Worcester Polytechnic Institute

Abstract

This Major Qualifying Project specialized in analyzing textual feedback messages for problems on the ASSISTments platform. Initially, work focused on designing and implementing a randomized control trial to measure the effectiveness of common wrong answer feedback messages. With work finished on the design of the RCT, focus shifted to generating a set of features using natural language processing techniques for each CWA feedback message, to be used after the completion of the RCT. After the generation of these features, the final stages of this project applied a similar method to hint messages in ASSISTments' TeacherASSIST system. NLP generated features were generated in order to analyze the effectiveness of hint messages, measured using next problem correctness as the dependent measure.

Acknowledgements

We would like to thank Professor Neil Heffernan for advising this project, which is a continuation of related work for the ongoing EIR project. This project would not have been possible without the continued guidance of Ashish Gurung and Russell Thompson, who were willing to help us learn as we learned more about experimental design, NLP models, and statistical analysis. Thank you to Ethan Prihar, for allowing us to build off of his work with the TeacherASSIST dataset. Finally, our thanks go out to the teachers and students who use the ASSISTments platform, who we hope to impact positively with the eventual results of this project.

Table of Contents

Abstract **2**

Acknowledgements **3**

1 Introduction **5**

2 Randomized Control Trial Groundwork **5**

2.1 CWA Feedback Messages 5

2.2 RCT design 6

2.3 Problem Set Creation 7

3 Feature Generation for CWA Feedback Messages **9**

3.1 Initial Steps 9

3.2 Parts of Speech Analysis 9

3.3 Word Choice Analysis 10

3.4 Future Work 11

4 Effectiveness of TeacherASSIST Hints **12**

4.1 Initial Steps 12

4.2 Parts of Speech Analysis 12

4.3 Word Choice Analysis 17

4.4 Future Work 18

Appendix - Parts of Speech **20**

1 Introduction

ASSISTments is a free homework and tutoring platform used by many school systems nationwide. As such, many different methods of student support have been created in an effort to improve the learning experience for the students using the platform. This project focused on textual feedback messages and hints within the ASSISTments platform, attempting to use natural language processing techniques to examine what features of a textual student support impact student learning in positive and negative ways. To do this, anonymized student data was used in conjunction with textual student supports and their associated metadata.

2 Randomized Control Trial Groundwork

2.1 CWA Feedback Messages

As part of the pre-existing EIR-CWA project, ASSISTments had collected data about common wrong answers to problems within the ASSISTments platform. A set of ASSISTments teacher-users had then been commissioned with writing feedback messages tailored to these common wrong answers (CWA feedback messages). Overall, feedback messages were written across problem sets within both the Illustrative Mathematics and Engage New York curricula. This project's initial task was to assist in the creation of an experimental design to test the effectiveness of these CWA feedback messages with respect to student learning.

2.2 RCT design

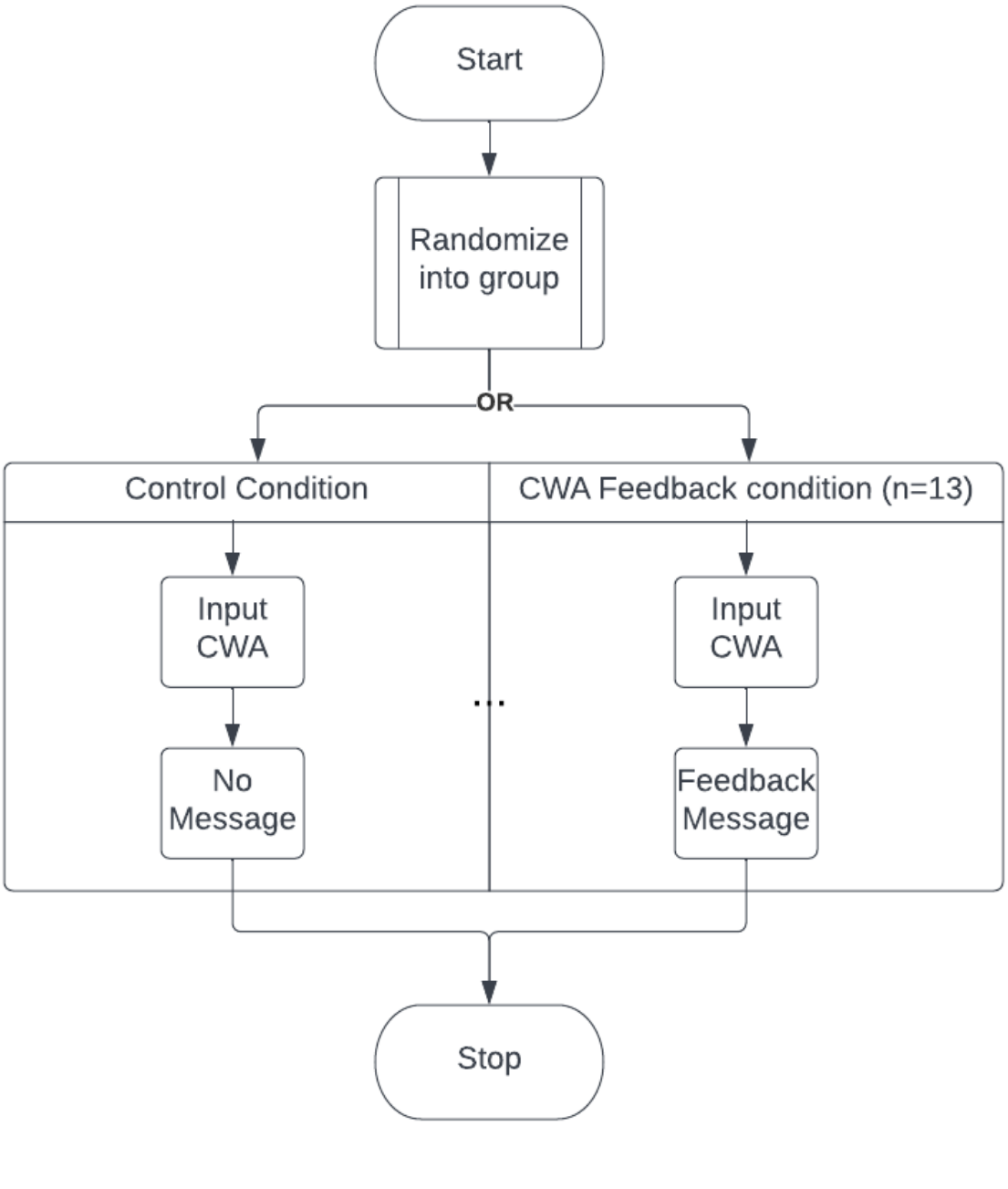

Figure 1: RCT Design over 14 Conditions (1 control, 13 experimental)

2.3 Problem Set Creation

Many experiments conducted through the ASSISTments platform rely on the collection of student data withinASSISTments 2.0. This poses a problem when trying to evaluate CWA feedback messages, since the delivery system for these messages is still in development. In order to start collecting data, the project team manually constructed 5 problem sets inside of the ASSISTments 1.0 problem set builder, using the pre-existing CWA feedback messages as mistake messages within the builder. The initial writing process for CWA feedback involved 12 teachers, with one additional teacher creating video-based CWA feedback messages. The problem sets created each had 14 sections, one for each teacher that created CWA feedback messages, and one control group that would not receive these messages.

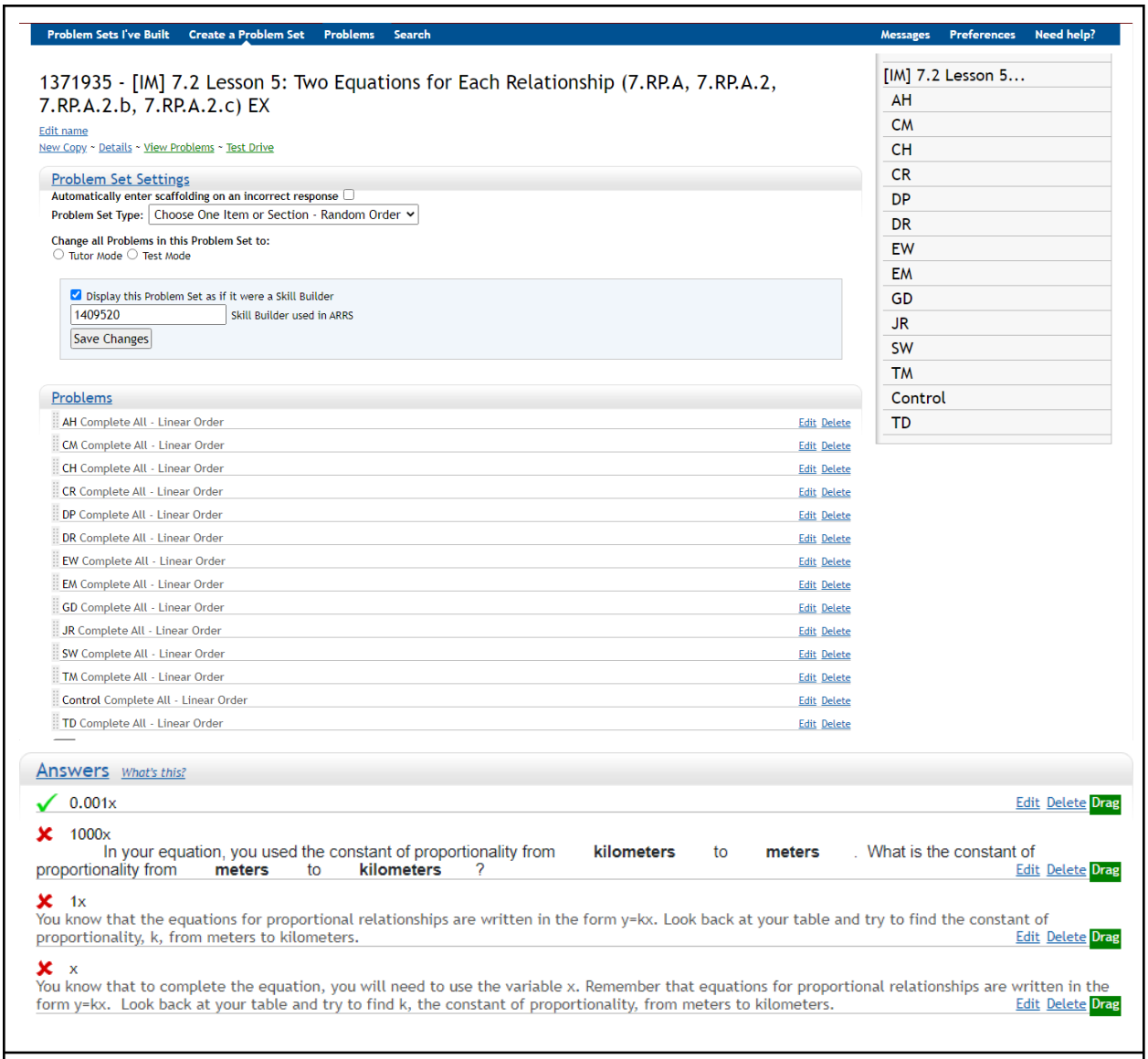

Figure 2: Sample Problem Set & CWA Feedback Messages

3 Feature Generation for CWA Feedback Messages

3.1 Initial Steps

As previously mentioned, the project team used the data from pre-existing CWA feedback messages. The data consisted of the feedback from a total of 13 teachers and a control group. In total, the uncleaned version of the dataset had around 7,500 rows, most of which were things that were harder to work with, and that we planned on omitting, specifically videos, things that were image-based, or open response questions. The first thing we did involved cleaning up the dataset. The main goal of our work here was to create features for a regression model that is to be run after the RCT is finished. The model will take a deeper look into what it takes to make a hint message be more effective using features created during this step.

3.2 Parts of Speech Analysis

Each CWA feedback message was tokenized using the spaCy NLP tokenizer to generate parts of speech counts for each feedback. This was not a trivial operation, generating these features required several intermediate steps. First, feedback messages must be stripped of HTML tags. After this stripping of tags, the spaCy tokenizer was applied, and part of speech tag counts were placed inside of a python Counter object. Finally, this counter object had to be unpacked and merged with the corresponding author's ID, along

with the problem the feedback was written for. The resulting dataset contains PoS counts for every CWA feedback within the ASSISTments database.

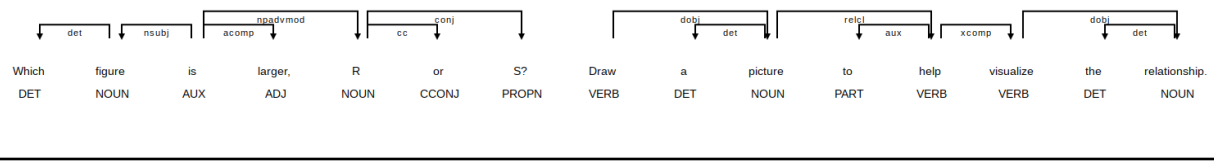

Figure 3: Parts of Speech Tagging for an Example CWA Feedback Message

3.3 Word Choice Analysis

While the parts of speech of the CWA feedback messages were being analyzed, we also took some time to look at the different kinds of words that the teachers used. We constructed a word cloud over all of the CWA feedback messages, shown in Figure 4, and also listed out the top 20 words over those messages. Then, we planned out 3 different categories to group the feedback into, and had words that would signify when a teacher's feedback message belonged to that group. The three groups we ended up going with were feedback messages that called upon knowledge that the student was expected to know prior to doing the problem, feedback messages that called upon knowledge that the current problem discussed, and feedback messages that specifically directed the answer the student gave. We ran this code that classified the different feedback messages over each of the teachers, and looked for trends in the teacher's writing styles.

Figure 4: Word Cloud over all of the CWA1 Feedback Messages

3.4 Future Work

The work with the CWA feedback was done so that we could take a deeper look into what makes a good feedback message. We plan on doing this after the randomized control trial finishes. The regression model we will use uses both the parts of speech analysis and the word choice analysis as features, which will be correlated with next problem correctness. Both the parts of speech and word choice analysis should provide enough variance to create a diverse enough feature set to have our model run off of.

4 Effectiveness of TeacherASSIST Hints

4.1 Initial Steps

After reaching a suitable point of progress working with CWA feedback messages, we then pivoted towards a similar form of analysis on a different, but similar, dataset: hint messages within the TeacherASSIST program. The dataset of hint messages we began exploring contained over 1.7 million rows, and was easily linked with next problem correctness. The goal of this portion of the project was to use the tools mentioned previously, part of speech tagging and word choice analysis, to generate features for these TeacherASSIST hint messages, to control for other variables which might impact a student's performance on their next problem, and examine any possible links between the features we generated and next problem correctness as measure of student learning.

4.2 Parts of Speech Analysis

Upon gaining access to the TeacherASSIST dataset, work immediately began on applying the methods used to analyze the CWA feedback messages to TeacherASSIST hints. A similar two-step process had to be applied to each hint message after stripping HTML tags, resulting in raw counts for each part of speech tag found in each hint message. There was an additional step which had to be taken after getting these counts, however. Depending on the problem, multiple hints exist, commonly designed to be seen in sequence in response to a student asking for a hint. For example, some problems have

three hints designed to be seen in a specific hint sequence. This posed a problem for conducting any analysis, since some students saw multiple hints for one problem. In order to make the student submission to hint message one-to-many rather than many-to-many, we exploited the knowledge that students will always be seeing these hints in the same sequence. If a student sees hint three for a given problem, they necessarily have seen hints one and two as well. This made it possible to have each row in our dataset function as an aggregate for the current hint and all hints that came before it in sequence, reflected in the part of speech counts.

SPACE	NUM	AUX	PUNCT	DET	NOUN	ADV	ADP	PROPN	VERB	PART	PRON	ADJ	SYM	SCONJ	CCONJ	INTJ	X	count
0.000000	0.230769	0.000000	0.076923	0.076923	0.384615	0.000000	0.000000	0.000000	0.076923	0.000000	0.000000	0.000000	0.153846	0.0	0.000000	0.0	0.0	13.0
0.000000	0.285714	0.000000	0.142857	0.047619	0.285714	0.000000	0.047619	0.000000	0.047619	0.000000	0.000000	0.000000	0.095238	0.0	0.047619	0.0	0.0	21.0
0.076923	0.115385	0.115385	0.115385	0.000000	0.153846	0.076923	0.076923	0.000000	0.076923	0.000000	0.115385	0.076923	0.000000	0.0	0.000000	0.0	0.0	26.0
0.071429	0.119048	0.095238	0.142857	0.023810	0.190476	0.071429	0.095238	0.000000	0.047619	0.000000	0.095238	0.047619	0.000000	0.0	0.000000	0.0	0.0	42.0
0.047619	0.000000	0.000000	0.095238	0.095238	0.190476	0.095238	0.095238	0.000000	0.190476	0.047619	0.047619	0.000000	0.000000	0.0	0.095238	0.0	0.0	21.0

Figure 5: Example Part of Speech Features After Normalization

After correcting the many-to-many relationship, further refinements to the part of speech features were made. First, the part of speech counts were translated into a normalized form representing the percentage of all words present in the given hint message that were the listed part of speech. Next, correlations between parts of speech were analyzed and considered, resulting in several of these parts of speech that often appeared together, or parts of speech serving similar functions, to be combined into larger categories.

DESC (Descriptor)	AUX, ADJ, ADV, PART
MATH (Mathematical)	SYM, NUM
SUBJ (Subject)	NOUN, PRON, PROPN
Unchanged	PUNCT, ADP, VERB, SCONJ, CCONJ, INTJ
Dropped	SPACE, DET, X

Table 1: Final Categories for each Part of Speech

For a detailed description of each part of speech present in the table above, please see the appendix.

Finally, additional factors that could have an impact on next problem correctness were considered and controlled for as part of the analysis. Two key factors were controlled for: student prior knowledge and problem difficulty. Prior knowledge was modeled using a student's average problem correctness, assuming a student with high average problem correctness is more likely to have prior knowledge. Problem difficulty was similarly modeled with problem average correctness under the assumption that problems with low average correctness are more difficult.

Variable	Coefficient	P > |t|	0.025	0.975
Intercept	0.3657	0.0	0.347	0.384
PUNCT	-0.3168	0.0	-0.349	-0.284
ADP	-0.1397	0.0	-0.176	-0.103
VERB	-0.2469	0.0	-0.247	-0.220
SCONJ	0.4727	0.0	0.410	0.536
CCONJ	-0.2096	0.0	-0.261	-0.158

INTJ	0.6437	0.243	-0.436	1.723
DESC	-0.1347	0.0	-0.158	-0.112
MATH	-0.2516	0.0	-0.280	-0.223
SUBJ	-0.1457	0.0	-0.174	-0.118
Problem Average Correctness	-0.0480	0.0	-0.054	-0.042
Student Average Correctness	0.2951	0.0	0.289	0.301

Table 2: Regression Results for Part of Speech Features, with $R^2 = 0.020$

Results from the regression analysis tentatively suggest subordinating conjunctions have a positive impact on the quality of a hint message, though user average correctness is still a dominant predictor for next problem correctness.

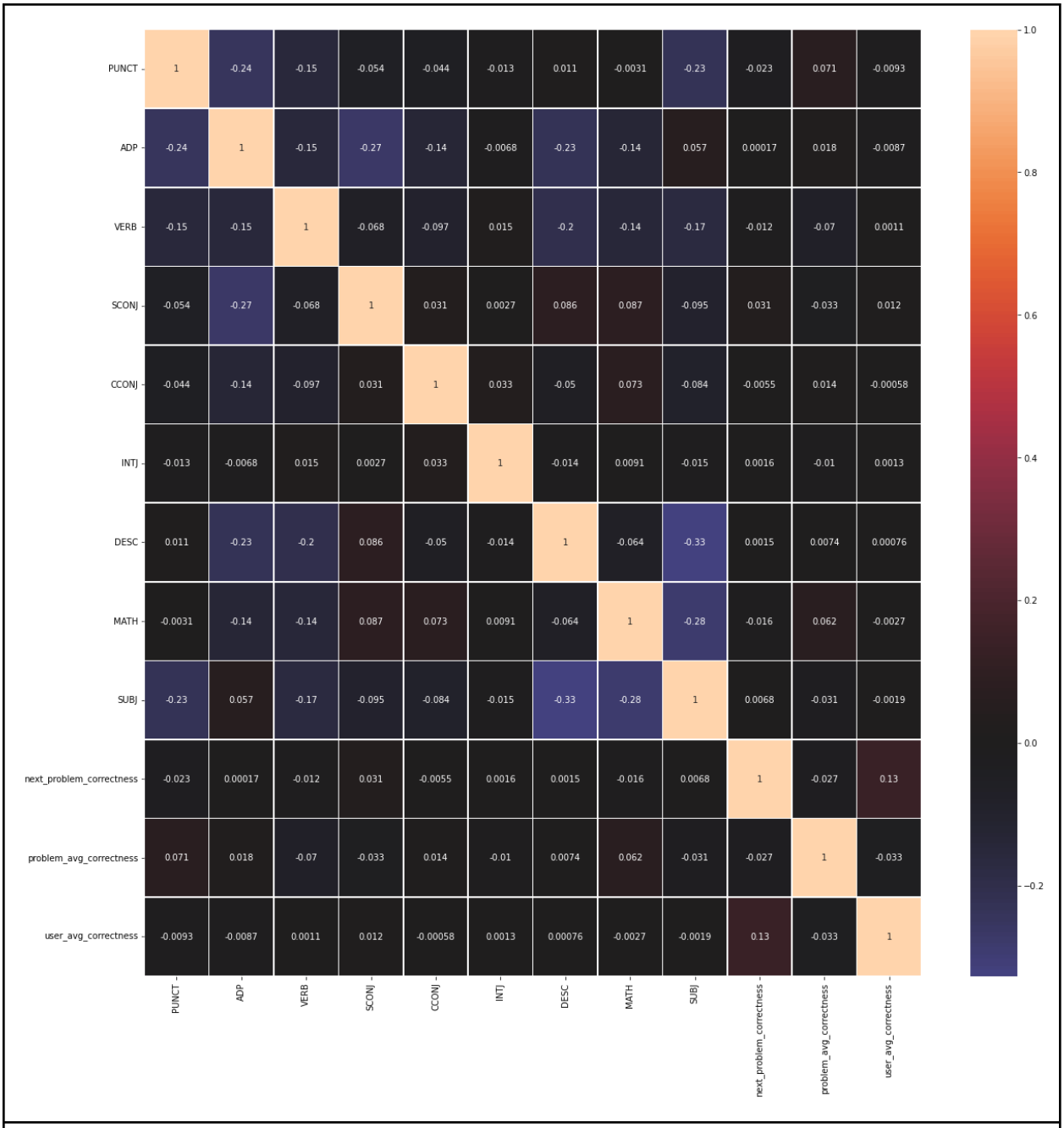

Figure 6: Correlation Heatmap for the Results Shown in Table 2

4.3 Word Choice Analysis

In addition to the parts of speech analysis on the TeacherASSIST dataset, we also conducted an analysis of the authors' word choice in different hint messages. We computed a list of the 50 most common words found in the dataset within the hint messages, and tied word choice to sentiment to create features related with the sentiment of the authors' hints. In the end, we ended up looking at whether the hint was in the form of a question, or if the hint contained any of the words what, if, how or could. We used these features to run a regression analysis to see how correlated each of those were to things like the current and next problem correctness, and how often the student asks for help or completes the problem set. Table 3, below, highlights the regression analysis that was run.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     next_problem_correctness   R-squared:                       0.068
Model:                                 OLS   Adj. R-squared:                  0.068
Method:                      Least Squares   F-statistic:                     5512.
Date:                     Wed, 23 Feb 2022   Prob (F-statistic):               0.00
Time:                             22:27:14   Log-Likelihood:            -5.3488e+05
No. Observations:                   833359   AIC:                         1.070e+06
Df Residuals:                       833347   BIC:                         1.070e+06
Df Model:                               11
Covariance Type:                 nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                        0.3183      0.007     46.206      0.000       0.305       0.332
iif                              0.0060      0.003      2.230      0.026       0.001       0.011
question                        -0.0331      0.002    -14.838      0.000      -0.038      -0.029
how                              0.0349      0.003     13.716      0.000       0.030       0.040
could                           -0.1478      0.023     -6.435      0.000      -0.193      -0.103
answer_given                    -0.1869      0.001   -138.301      0.000      -0.190      -0.184
problem_hint_count              -0.0118      0.001    -19.549      0.000      -0.013      -0.011
problem_avg_correctness         -0.1295      0.003    -50.045      0.000      -0.135      -0.124
user_avg_correctness             0.1883      0.004     48.373      0.000       0.181       0.196
next_problem_avg_correctness     0.4292      0.003    160.181      0.000       0.424       0.434
user_avg_support_requested      -0.1226      0.005    -22.302      0.000      -0.133      -0.112
user_avg_completed              -0.0549      0.007     -7.610      0.000      -0.069      -0.041
==============================================================================
Omnibus:                 10017176.798   Durbin-Watson:                   0.232
Prob(Omnibus):                 0.000   Jarque-Bera (JB):           111592.611
Skew:                          0.577   Prob(JB):                         0.00
Kurtosis:                      1.627   Cond. No.                         130.
==============================================================================
```

Table 3: Regression Analysis ran on word choice in TeacherASSIST dataset

4.4 Future Work

Unfortunately, shortly before the conclusion of this project, it was discovered that there was an error in the initial TeacherASSIST dataset that impacts the applicability of any correlations we found. Duplicate hint messages existed as part of the dataset, with several entries for several hints where authors had gone back and edited their prior work. This means that establishing a proper hint to hint comparison for statistical analysis is virtually impossible at this time. However, the techniques used to generate features for these hint messages remain promising. Assuming an updated version of this

dataset is created without duplicate entries, it would take a minimal amount of work to use the work from this MQP in order to construct new features for the updated dataset. Additionally, TeacherASSIST also contains within it textual explanations for many problems on the ASSISTments platform. A similar exploration of explanations and their effects on knowledge retention utilizing natural language processing techniques to identify features of effective explanations should be undertaken. Future work should attempt to analyze the quality of hints and feedback messages specifically for low-knowledge students.

Appendix - Parts of Speech

Tag	Full Name	Description	Example
ADJ	Adjective	Modifies a noun	Younger, oldest
ADP	Adposition	Prepositions & Postpositions	In, to, before
ADV	Adverb	Modifies a verb	Very, quickly, never, exactly
AUX	Auxiliary	Often specifies tense of a verb	Has, will, should
CCONJ	Coordinating Conjunction	Links two clauses without subordination	And, or, but
DET	Determiner	Often specifies or generalizes a noun	This, which, a, the
INTJ	Interjection	Expresses an emotional exclamation	Great, ouch, hey
NOUN	Noun	Subject, denotes person, place, thing, or idea	Boy, dog, flower, love
NUM	Numeral	Expresses a number	1, 90, four, XXIV
PART	Particle	Associated with another word in order to enhance meaning	‘s (possessive marker), not
PRON	Pronoun	Word that substitutes for a noun	I, you, they, nobody,
PROPN	Proper Noun	Denotes a specific noun	John F. Kennedy, Worcester, WPI
SCONJ	Subordinating Conjunction	Links two clauses, implying one clause depends on the other	If/then, while, that
SPACE	Space	A space character	‘ ‘
SYM	Symbol	Non-alphanumeric symbols	@, +, =, >
VERB	Verb	Action words, done by a noun	Read,, sitting, sat
X	Other	Word not recognized as another part of speech	Skldh, nvkdsjaoui