Assessing English Essay Proficiency Scores

W207 Final Project
Dec 2022

Stephanie He, Sarah Hoover, Joseph Roberts, Stephen Tan

Overview

- Motivation and History
- Dataset and Exploratory Data Analysis
- Approach and Model Experiments
- Model Outcomes
- Conclusion and Future Steps

Github Repo: https://github.com/joethequant/kaggle_english_language_grading

Competition: https://www.kaggle.com/competitions/feedback-prize-english-language-learning

Motivation

Stephanie

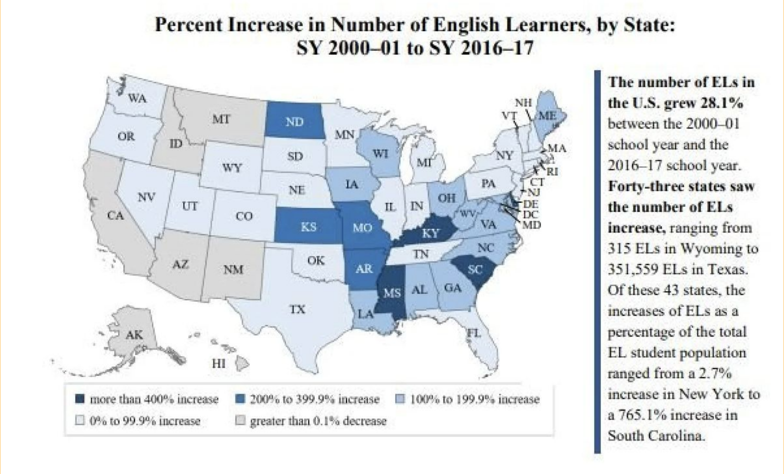
Motivation

Research questions:

- Use ML models to predict English Proficiency Score of 8th - 12th grade student essays

Solve education problem

- **Rapid growing English Language Learners ("ELLs")** population
- **Limited teacher resources** to provide timely feedback so students are aces lack of practice
- **Limitation on Assisted Writing Feedback tools** (AWFTs)
- **Using ML model to provide more accurate feedback** to further improve writing



What Have Been Done?

Numerous models of Automatic Essay Grading ("AEG") have been developed since the 1960s*

- **PEG (Project Essay Grade) - 1966**
  - Multiple regression - using a number of easily quantifiable variables
- **IEA (Intelligent Essay Assessor ("IEA") - late 1990s**
  - Use TF-IDF (Frequency and inverse document frequency) to drive text-word matrix
  - Use LSA (Latent Semantic Analysis)
- **E-rater - late 1990s**
  - Based on NLP with AL and a regression algorithm
  - Three main natural language processing tools: Syntactic, expository and thematic analysis
- **BETSY (Bayesian Essay Test Scoring System) - open source and free**
  - Integrate content and formal features into one feature set and classified the scores into four levels
  - Use multivariate Bernoulli model and Bernoulli model

Data

Stephen

Dataset

	text_id	full_text	cohesion	syntax	vocabulary	phraseology	grammar	conventions
2149	A0B2BF94231C	First impressions are almost impossible to change. I disagree because when you look at a person for less than 6 se However the way a person looks from your first impression isnt always who they really are. Your observation on so Personality can change your first impression. Personality is mostly who you are finding out, who that person is, car Also communication is key its really important to have communication. Communication gives confidence to the oth Additionally, i disagree that the first impression is almost impossible to change. Their personality, the way they look	4.0	3.5	3.5	4.0	3.5	4.0
953	48F7FCAD8B23	Will has you can see getting advice from other's is not a bad idea. I think getting advice from others has a great po My first statement is that how ever is trying to get advice should get it form multiple people not just from one perso My second statement is that getting advice form people can have a good or bad point of view but it also can depe My last statement is that asking people that you thing that will tell you the truth no matter if it hurts or not is not a b In conclusion, asking for advice sometimes and talking to others more advice can help you make good choice but In this world you should take most of the advice that other people gave you because they have already been thoug	3.0	3.0	3.5	3.0	2.0	2.0
2315	AC8331539332	"Do we choose our own character traits, or our character formed by influences beyond our control?" I think you sh You shouldn't be acting like someone your not, a lot of people act certain ways or try to be someone they're, usual My point in all of this is just be yourself, trust yourself, have confidence what you wanna be and don't let somebod But, I agree that you should be your own character , because you can't trust everybody or expect everybody to ha	2.5	3.5	3.5	3.5	2.5	3.0
2026	98D8FF9E3C56	The life is being more modern, people also need to improve themselves. That's why some schools want to offer so This plan has two positions,some people agree, but others don't think it's a good idea. Is it a good idea? It is. It is a Let's start with the first reason now. Do you agree that the time is limit? The time is the only thing you can't move b The next reason will be the reason help students the most. Most students who follow this plan will have a lot of exp The last reason refers to elective courses. Many students like these courses, but some students don't. These progr On the other hand, many students do not agree this plan. They think it gives stressful for them, they worry about th Conclusion, we should support some schools offer these programs that allow high school students to graduate ear	4.5	3.5	4.0	3.5	3.0	3.5

*__Definitions__: _Cohesion_: unity and connectedness; _Syntax_: word arrangement; _Vocabulary_: words used; _Phraseology_: expressions (idioms and phrases); _Grammar_: structural constraints; _Conventions_: writing mechanics (capitalization and punctuation)

Dataset

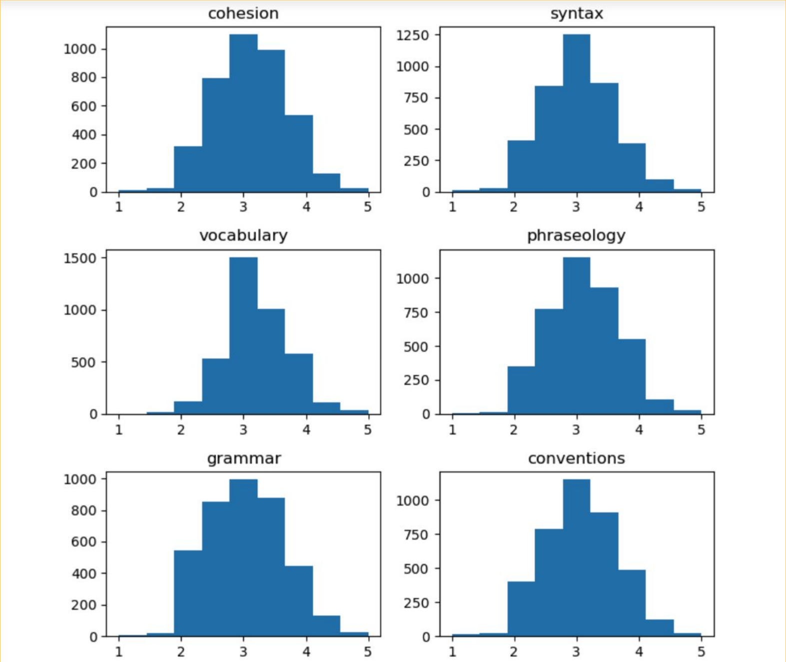
Sourced from Kaggle

- Scored written (English) essays by 8th - 12th graders
- Predetermined train/test split
    - Training set: 3,911 unique essays
        - We did a 70/10/20 split for training/validation/test
    - Testing set: 3 unique essays
        - ~2,700 hidden essays
- Feature - Entire essay
- Output* (6) - Scored 1.0 to 5.0 with 0.5 increments
    - Cohesion, Syntax, Vocabulary, Phraseology, Grammar, Conventions

**Definitions*: Cohesion: unity and connectedness; Syntax: word arrangement; Vocabulary: words used; Phraseology: expressions (idioms and phrases); Grammar: structural constraints; Conventions: writing mechanics (capitalization and punctuation)*

Dataset EDA

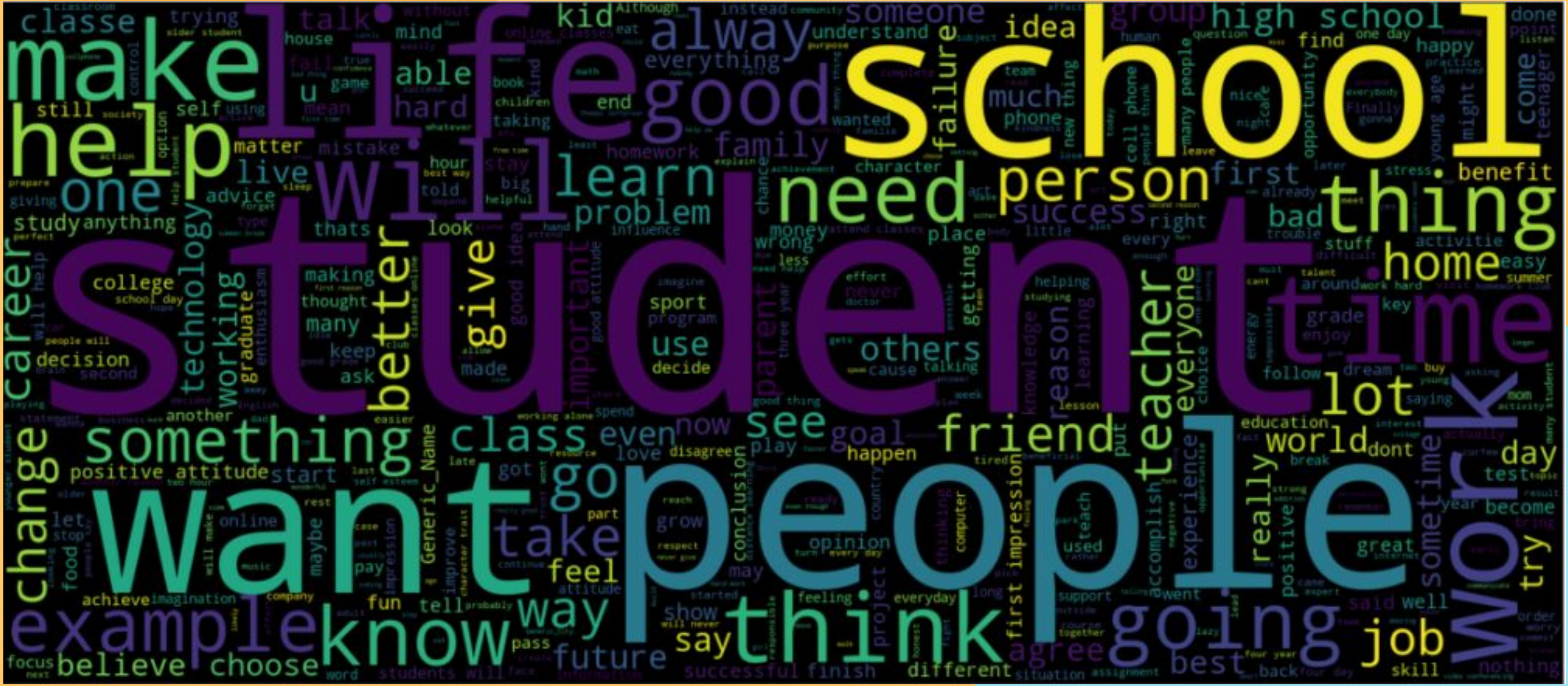
Score distribution



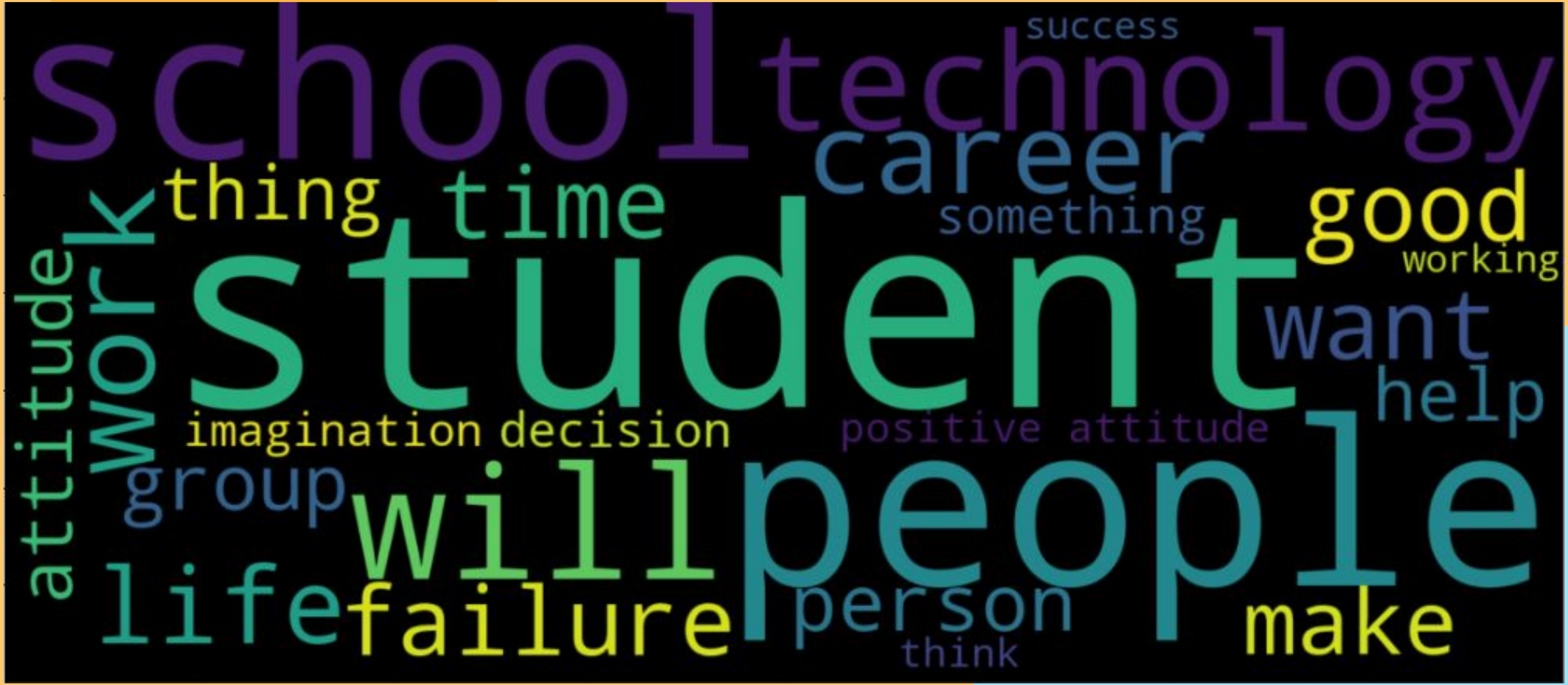
Correlation matrix

**Definitions*: <u>Cohesion</u>:unity and connectedness; <u>Syntax</u>: word arrangement; <u>Vocabulary</u>: words used; <u>Phraseology</u>: expressions (idioms and phrases); <u>Grammar</u>: structural constraints; <u>Conventions</u>: writing mechanics (capitalization and punctuation)*

Top 500 Most Used Words



Top 25 Most Used Words



Additional Feature Creation

word count (essay length)

average =
430 words

sentence count

average =
18 sentences

sentence length

average =
30 words

word length

average =
4 letters

repeated words

average =
35 words

stop words

average =
47 words

spelling errors

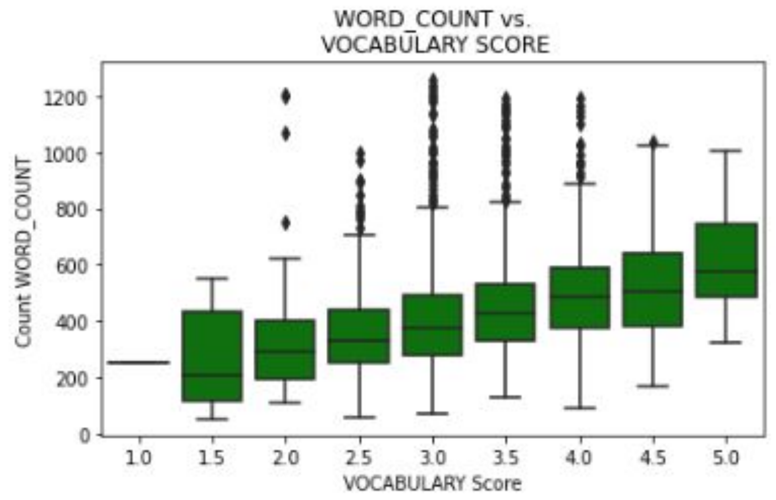
average =
4 words

Visualize New Features by Measure Scores

Box and whisker plots

- Understand distribution
- Identify outliers

Plots based on
2,815 out of 3,811 essays
(training dataset)



Scores (x-axis) increase from left to right

MEASURES

cohesion | syntax | vocab | phraseology | grammar | cohesion

NEW FEATURES

word count

word length

sentence count

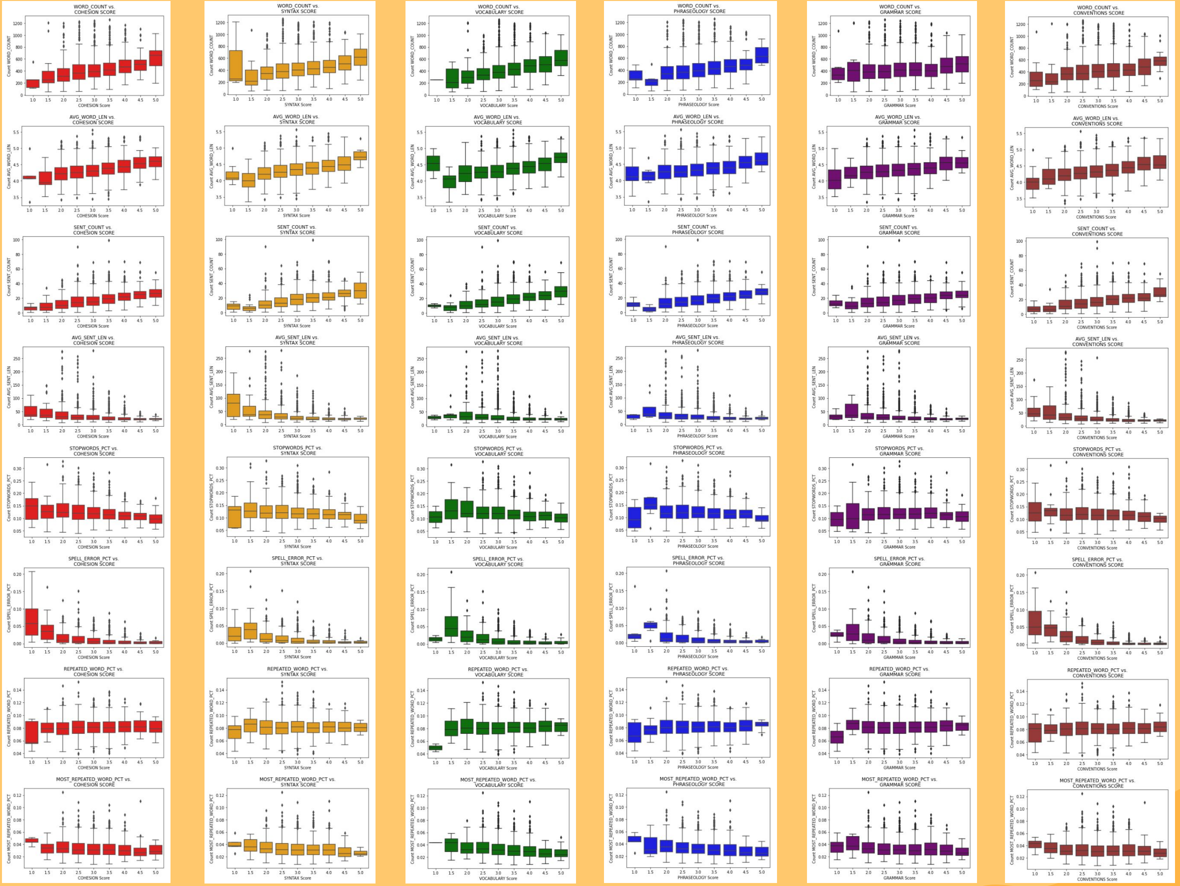
sentence length

stopwords %

spelling error %

repeated word %

most repeated word %



Approach and Experiments

Joe

Algorithm Evaluation - Continuous

Evaluation Function: Root Mean Squared Error

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{ij} - \hat{y}_{ij}\right)^2}$$

$$\begin{aligned} N_t &= \text{number of predicted variables} \\ n &= \text{number of test samples} \\ y_{ij} &= i\text{-th observed value of } j\text{-th variable} \\ \hat{y}_{ij} &= i\text{-th predicted value of } j\text{-th variable} \end{aligned}$$

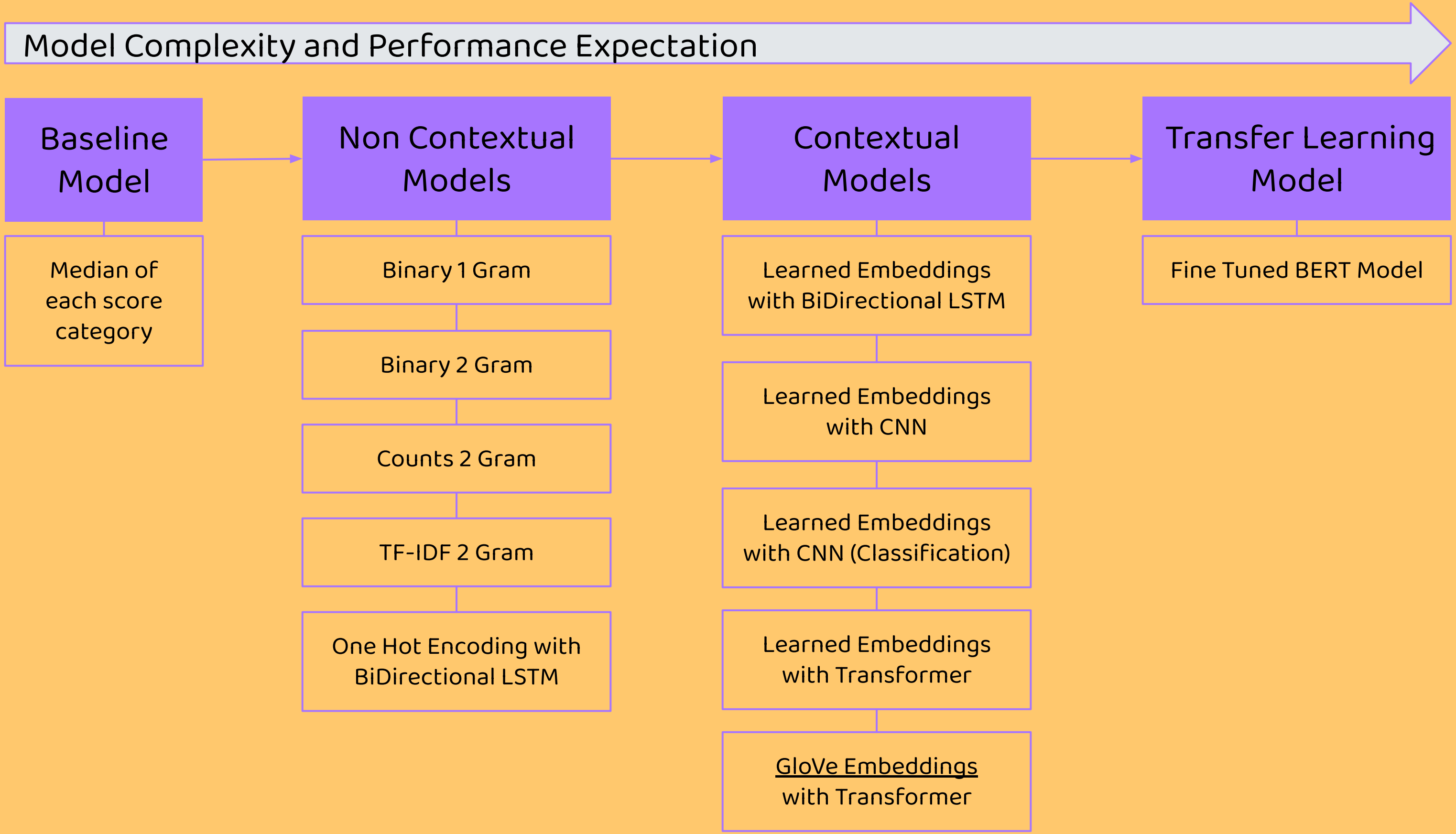
Loss Function: Mean Squared Error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

$$\begin{aligned} n &= \text{number of test samples} \\ y_i &= i\text{-th observed value} \\ \hat{y}_i &= i\text{-th predicted value} \end{aligned}$$

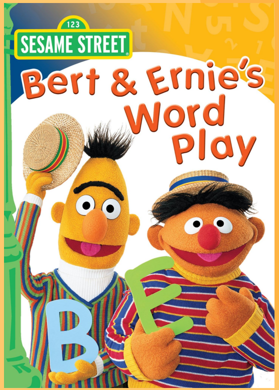
LOWER RMSE IS BETTER!!!!

Approach


Model Complexity and Performance Expectation
Baseline Model
Median of each score category
Non Contextual Models
Binary 1 Gram
Binary 2 Gram
Counts 2 Gram
TF-IDF 2 Gram
One Hot Encoding with BiDirectional LSTM
Contextual Models
Learned Embeddings with BiDirectional LSTM
Learned Embeddings with CNN
Learned Embeddings with CNN (Classification)
Learned Embeddings with Transformer
GloVe Embeddings with Transformer
Transfer Learning Model
Fine Tuned BERT Model

Experiments

- The BERT Transfer Learning Model performed the best out of all optimized algorithms.
- We saw significant improvements with each larger architecture change (baseline ⇒ learned embeddings ⇒ BERT transfer model)
- The winning algorithm used a RoBERTa transfer model with a RMSE score of .43



Models	Test RMSE	Competition RMSE
Baseline - Median score of training data	.66	.65
Binary 1 Gram Tokens (20k Max)	.62	.61
Binary 2 Gram Tokens (20k Max)	.61	.60
Count 2 Gram Tokens (20k Max)	.66	.65
TF-IDF Norm with 2 Gram (20k Max)	.66	
One Hot Encoding with BiDirectional LSTM	.62	.59

Models	Test RMSE	Competition RMSE
Embedding with BiDirectional LSTM	.64	.61
Embedding with CNN	.59	.57
Embeddings with Transformer	.58	.57
GloVe Embeddings with Transformer	.59	
BERT Transfer Learning Model Trainable	.51	.49

LOWER RMSE IS BETTER!!!!

Diving into Our BERT Model

Essays (1536 Max Tokens)

Part 1 | Part 2 | Part 3

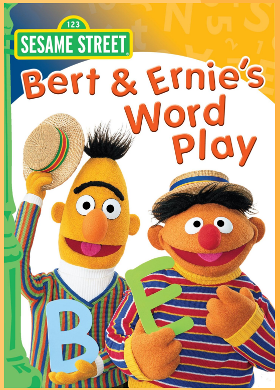
BERT

Mean Pooling | Mean Pooling | Mean Pooling

Concatenate

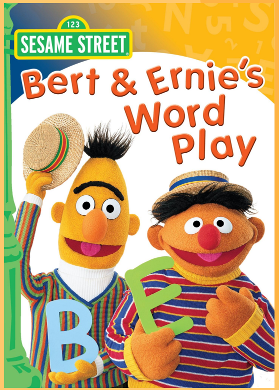
Hidden Layers (768, 512, 256 Dim)

Dense Layer (6)



- BERT only allows 512 tokens max
  ⇒ Split essays into 3 inputs
- The target variable is 1.0 through 5.0 in increments of 0.5 and in the competition we were scored on RMSE, thus we treated the target variable as continuous.
- Normalizing the target variable did not make a difference in model training.

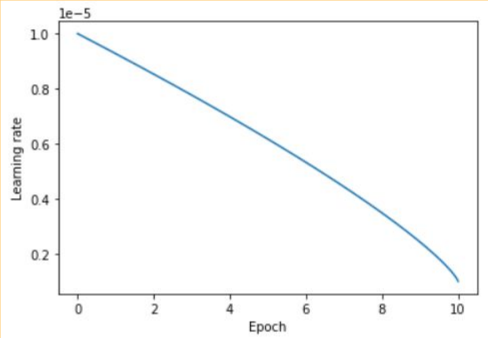
Diving into Our BERT Model



Optimization

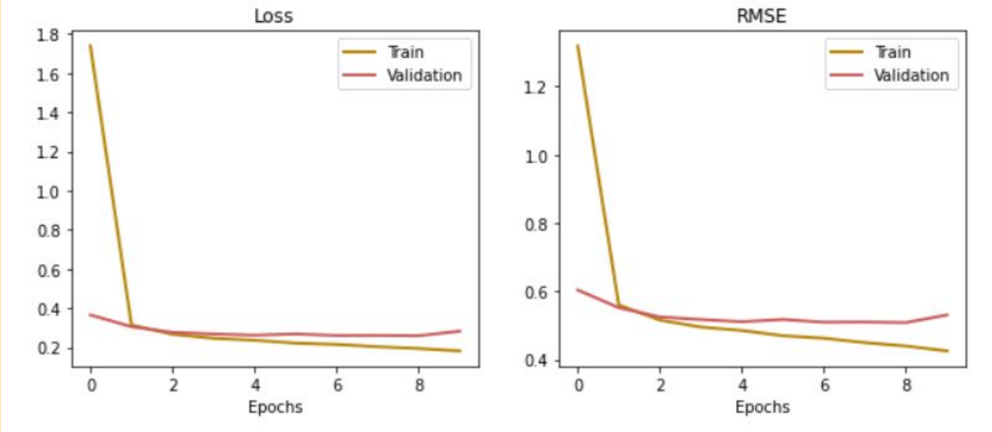
Algo: ADAM

Learning Rate: Decaying Learning Rate with a .8 power Polynomial starting at 1e-5 and ending at 1e-8. BERT used an ADAMw or weighted ADAM learning rate with warm up epochs.



Learning

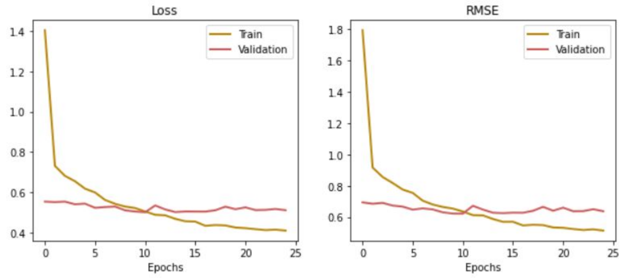
- We see the model perform relatively well compared to the other models almost immediately.
- In this model we fine tuned it by allowing the training to adjust the BERT model, but we saw little difference in final performance.



Outcomes

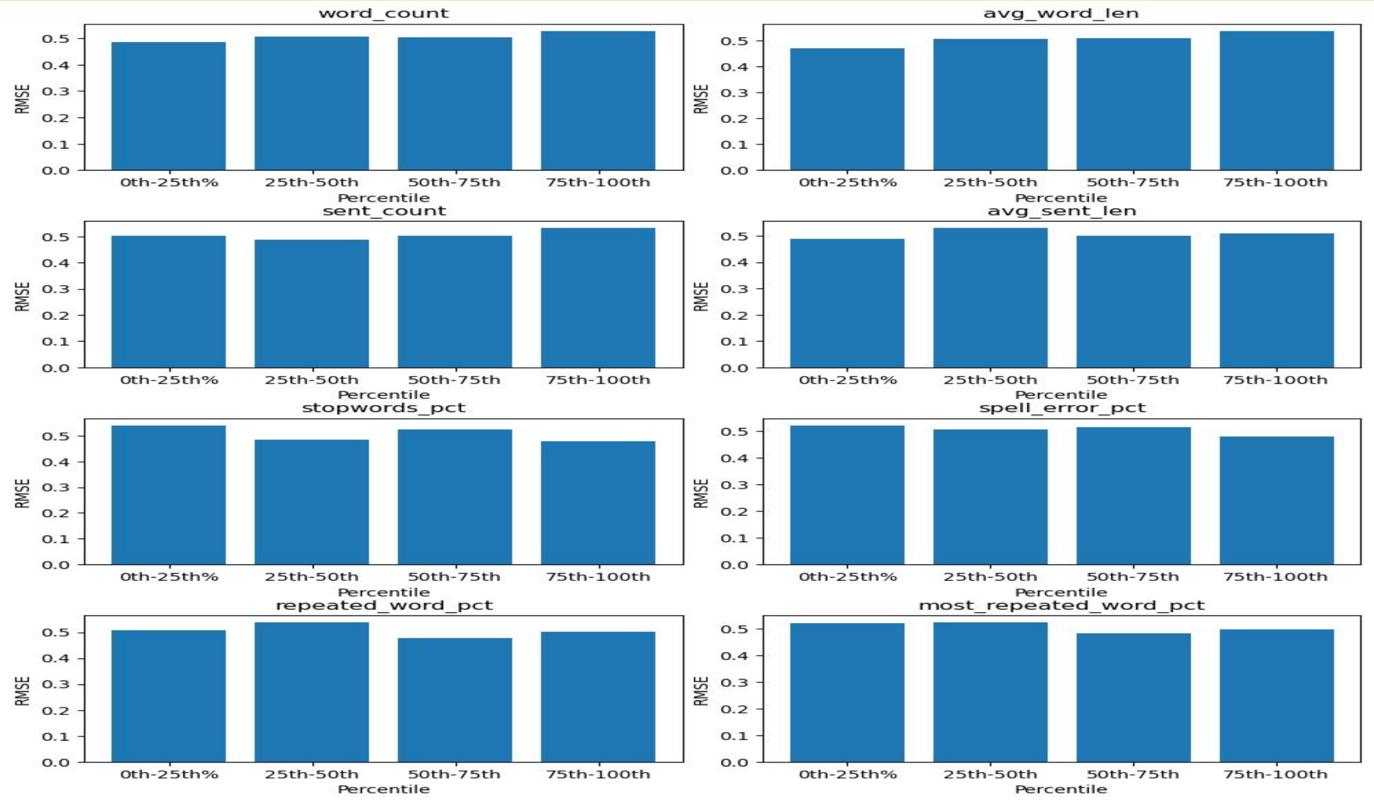
Sarah

Hyperparameter tuning for BERT model

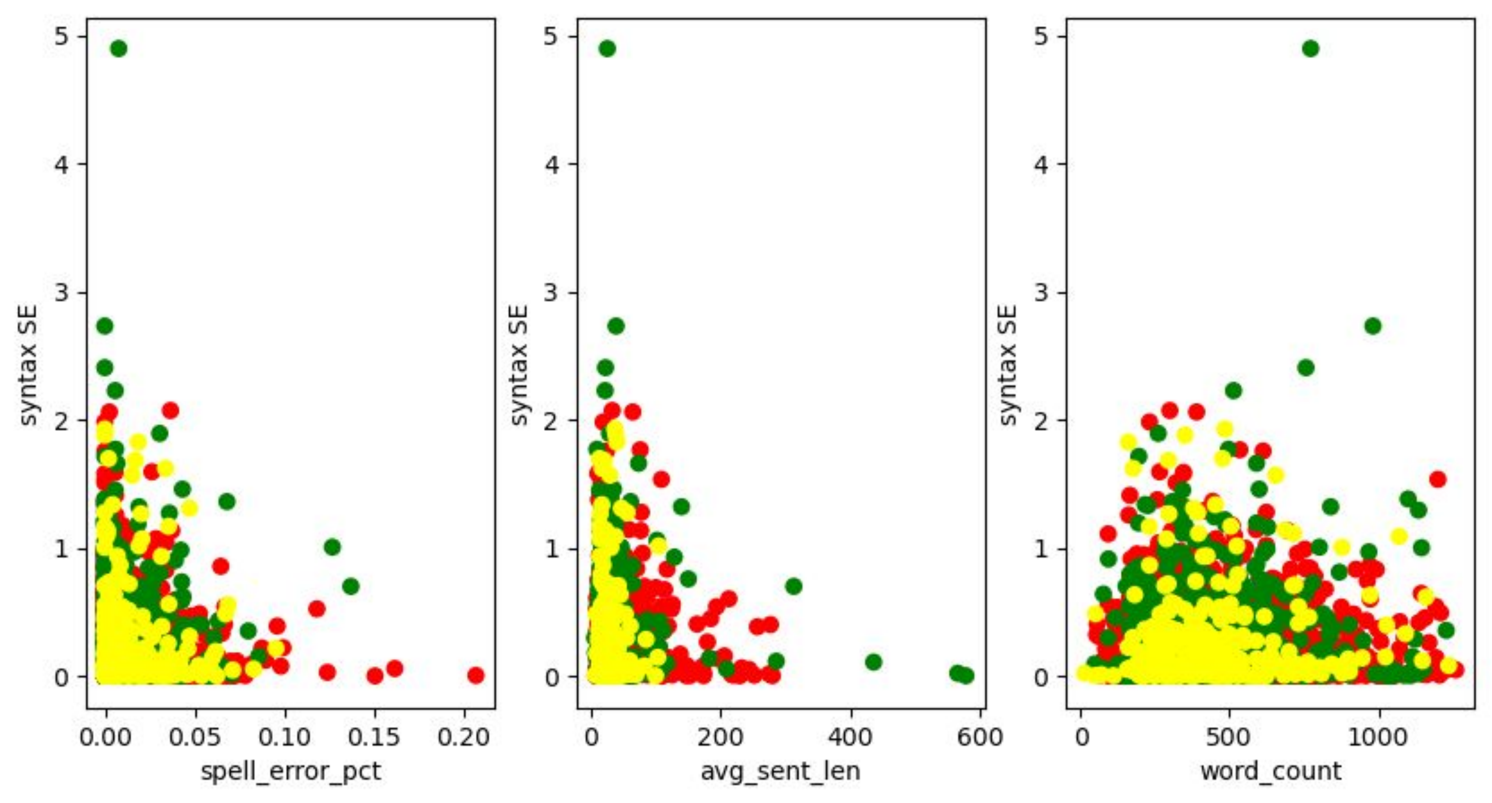


- Number of epochs
  - Started with 25
  - After viewing graphs, stopped at 10 epochs
- Max essay length
  - BERT max. 512 tokens
  - Breaking essays into chunks of 512 tokens and averaging the outcome did little to improve the predictions
- Freeze vs. Unfreeze embeddings
  - Surprisingly, BERT embeddings performed well for our task as-is

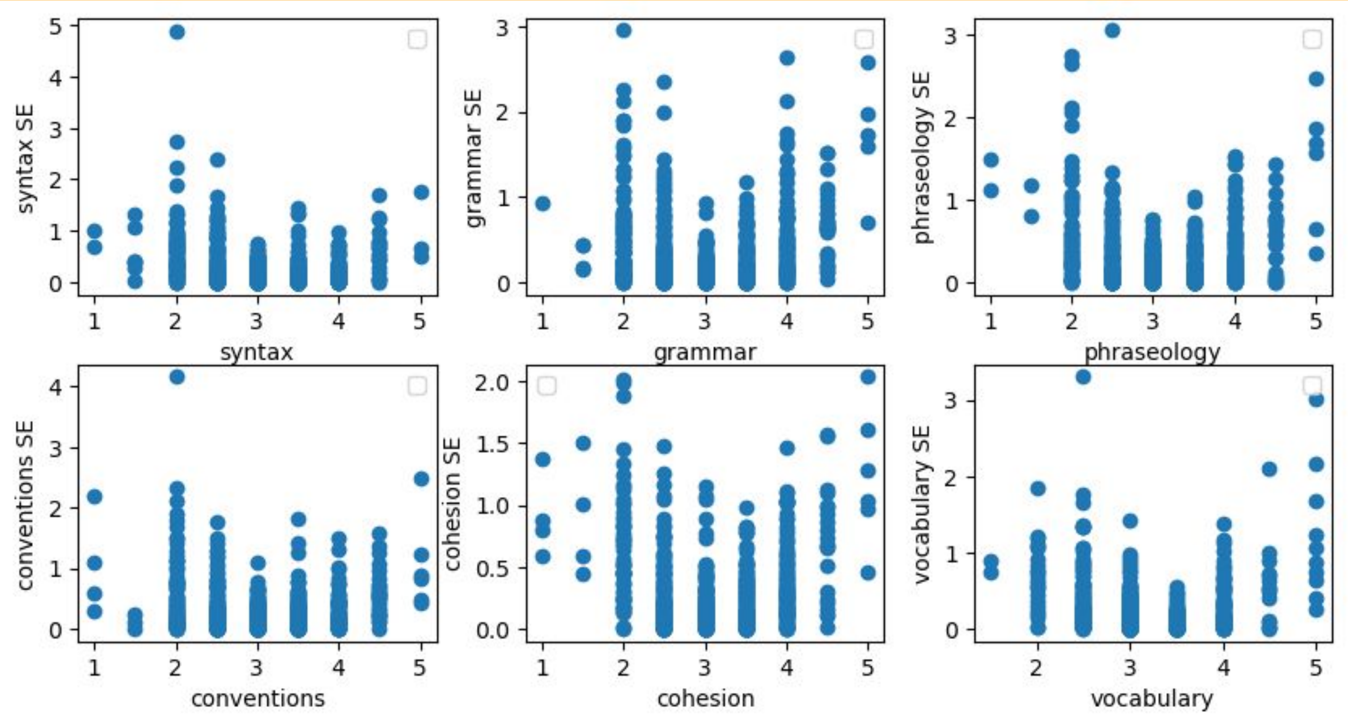
Subgroup analysis



Feature Correlation



Outcome comparison



Conclusions

Stephanie

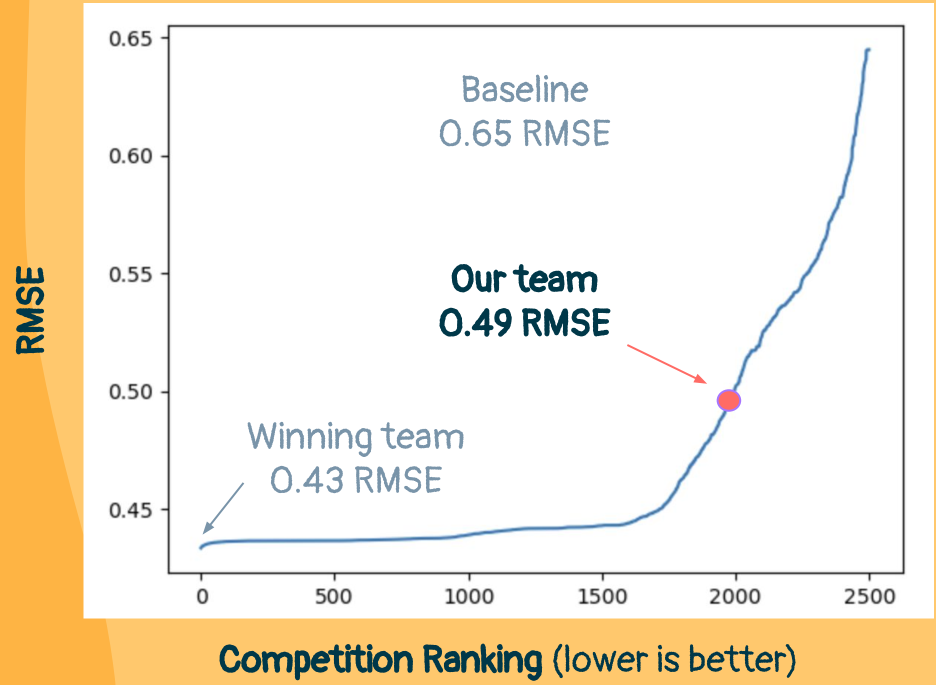
Conclusions

Key takeaways:

- **BERT transfer learning model is the best so far**, significant improve the results, because of its huge information on human language
- ...but computational heavy
- Pre-trained architecture embedding is important
- Use max 512 tokens or breaking the essay into three parts (the longest essay has ~1400 words) does not significantly change the results
- Newly engineered features does not provide much additional information because BERT embedding should have incorporated the information

Conclusions

Kaggle Competition Final Result (RMSE)



Kaggle Competition Results:

- 2654 teams joined
- We placed around 2000th

Future work:

- Use different forms of DeBERTa and RoBERTa model and pretrained embeddings
- Other techniques worked include:
  - Different pooling techniques
  - Different max_len
  - Pseudo labels

Contributions

Stephanie He: feature engineering, field research, subgroup analysis, slides

Sarah Hoover: summary statistics of outcomes, learned embeddings model with logistic regression, subgroup analysis and feature correlation for BERT model, slides on model outcomes

Joseph Roberts: Github management, data analysis, modeling, slides

Stephen Tan: EDA, feature engineering, box-and-whisker plots, slides + formatting

Thanks!

References

Fuzhuang Zhang, Lan Yu, and Jun Shen, "Automatic Scoring of English Essays Based on Machine Learning Technology in a Wireless Network Environment", May 2022 *https://www.hindawi.com/journals/scn/2022/9336298/*

https://arxiv.org/abs/2206.08232

https://towardsdatascience.com/effectively-pre-processing-the-text-data-part-1-text-cleaning-9ecae119cb3e

https://machinelearningmastery.com/clean-text-machine-learning-python/

https://www.hindawi.com/journals/cin/2021/8545686/

https://aclanthology.org/2020.pam-1.15.pdf

https://www.wwp.northeastern.edu/lab/wwvt/resources/introduction/index.html

https://www.geeksforgeeks.org/string-punctuation-in-python/