Machine Learning Final Project Proposal

Peizhong Gao
2023213526
Academic of Arts & Design
Tsinghua University
gpz23@mails.tsinghua.edu.cn

Yuzhe Zhang
2023215229
Global Innovation Exchange
Tsinghua University
zhang-yz23@mails.tsinghua.edu.cn

Matthew Taruno
2023280596
Global Innovation Exchange
Tsinghua University
wujiahui23@mails.tsinghua.edu.cn

1 Background

Writing is a fundamental to your success and self-improvement. From just *the way you type and move your mouse*, could we predict how well you will score on a writing test? These insights might help you understand what writing behaviors lead to better writing performance. To collect data on the way humans type and move their mouse during a writing task, the exact environment works as follows. Participants of this project were hired from a crowdsourcing platform called Amazon Mechanical Turk to do an argumentative writing task. 2471 participants were asked to write an argumentative essay within 30 minutes in response to a writing prompt adapted from a retired SAT test taken by high school students. While participants took this writing test, a keystroke logging program written in JavaScript was embedded into the test-taking platform website. These events (median of 3000) fundamentally include when the key/mouse was pressed and released, action time, cursor positions, type of keyboard action, and the actual words typed.

At first, this may seem unintuitive. In the SAT grading scheme, there are usually two scorers who read each essay independently and provide a score of 1 to 4 (competition data ranges from 1 to 6), graded on dimensions such as comprehension of essay reading passages, analysis and explanation of argument, and effectiveness of essay writing. While this model might directly help graders with grading writing tests, we believe this quantitative insight is most useful for understanding the psychological processes behind writing, nature of these writing tests that millions of children have to do every year, and maybe even insights on how you could be a better writer.

2 Definition

Each test is graded on an essay score received out of 6 (the prediction target for the competition). This target variable of test score is what we want to predict, and the competition is evaluated using the RMSE: $RMSE = (\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2)^{1/2}$

Alternatively, there is a prize that rewards model efficiency because highly accurate models tend to be computationally heavy and leave a stronger carbon footprint. These models would like to be used to help educational organizations with limited computational capacities, so here both runtime and predictive performance is measured: $Efficiency = \frac{RMSE}{Base-minRMSE} + \frac{RuntimeSeconds}{32400}$

Preprint. Under review.

3 Related Work

Keystroke analysis has been increasingly used to gain insight into the writing process. With keystroke logging, information on every key pressed and released is recorded, resulting in a detailed log on students' text composition. [5, 8]

Drawing from existing scholarly works, we categorize these features into five distinct groups: (1) Pause-related features encompass aspects like intervals between keystrokes and pauses preceding word initiation, as highlighted in Barkaoui's 2016 study [2] and in Medimorec & Risko study [9]; (2) Revision-oriented features, which include metrics such as the frequency of backspace usage and the duration of editing pauses[4]; (3) Verbosity-related features, chiefly represented by the total word count [7]; (4) Fluency-related features, exemplified by the proportion of typing bursts culminating in revisions[1]; and (5) Features pertaining to non-character-producing keystrokes, encompassing actions like text selection, copy-paste operations, cut commands, and mouse navigations. These categorizations aid in a comprehensive understanding and analysis of typing behaviors and patterns.[6]

Machine learning method was introduced to predict final grade and classify students who might need support at several points during the writing process but results are even worse than baselines, which seem to point out that the relationship between keystroke data and writing quality might be less clear than previously posited. [3]

4 Proposed Method

Feature Engineering The baseline features engineered explained in the competition involve rate of written language production, pauses, revisions, and writing session bursts of no revisions or deletions. In terms of the actual written text, for the purposes of the competition, the data is anonymized as q so we are not able to predict the writing score based on the essay prompt and writing content. However, we still are able to analyze the sentence structure, word length, and difficulty. The sentence may look something like this: ”qqqqqqqqq qq qqqq qqq q...” From here, it's the philosophical question of predicting $P(score|graderpreferences)$ from subset of human population of graders. For examples, graders might prefer essays with a higher frequency of ”difficult” words, which we can infer by the number of characters of each word. It could be worth exploring more about who this population is. Additionally, as can be seen in the ”Related Works” section, we will use some psychological background and domain knowledge to hopefully engineering highly-predictive features.

Definite Approaches: (1) Ensemble models (combining predictions from various models) with techniques like bagging, boosting, stacking, and blending to leverage the strengths of each model to improve the overall prediction. (2) Persistent Experimentation. We will experiment with a variety of approaches, namely the combinations of many models and feature sets to find the winning combination. Additionally, we may use cross validation and feature selection to remove non-informative predictors. (3) Interpretability. Perhaps we can use SHAP or LIME to interpret the model and understand which features are most predictive of writing quality.

Exploratory Ideas: (1) External Data and Data Augmentation. Adding more data examples - perhaps through finding relevant datasets and using pseudo labelled data - will be considered. (2) Evaluation Metric. At the very baseline, we should make sure our model's evaluation metric mirror's the competition's evaluation metric as close as possible. Post processing could be added to the data, since this competition uses MSE, we can consider clipping predictions to prevent extreme values from having a disproportionate effect. Or use ensemble residual correction. (3) Transfer learning for Behavioral Pattern Recognition. If there exists a model pretrained on keystroke dynamics for other purposes, such as user authentication or emotion detection based on typing patterns, this could serve as a starting point. The learned representations of typing behavior might transfer to the task of predicting writing quality, capturing subtle aspects of writing fluency and process.

Model Efficiency: Since the competition also focuses on model efficiency, we plan to employ a strategic mix of lightweight yet powerful models, complemented by robust feature engineering to enhance predictive power without increasing complexity. Dimensionality reduction techniques can help streamline input features, while model pruning and quantization can reduce computational load. Knowledge distillation can transfer insights from complex models to more compact ones. Multithreading may help as well, abiding by the constraints of the competition.

References

[1] Veerle M Baaijen, David Galbraith, and Kees De Glopper. Keystroke analysis: Reflections on procedures and measures. *Written Communication*, 29(3):246–277, 2012.

[2] Khaled Barkaoui. What and when second-language learners revise when responding to timed writing tasks on the computer: The roles of task type, second language proficiency, and keyboarding skills. *The Modern Language Journal*, 100(1):320–340, 2016.

[3] Rianne Conijn, Christine Cook, Menno van Zaanen, and Luuk Van Waes. Early prediction of writing quality using keystroke logging. *International Journal of Artificial Intelligence in Education*, 32(4):835–866, 2022.

[4] Paul Deane. Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks. *ETS Research Report Series*, 2014(1):1–23, 2014.

[5] Mariëlle Leijten and Luuk Van Waes. Keystroke logging in writing research: Using inputlog to analyze and visualize writing processes. *Written Communication*, 30(3):358–392, 2013.

[6] Mariëlle Leijten, Luuk Van Waes, Iris Schrijver, Sarah Bernolet, and Lieve Vangehuchten. Mapping master's students'use of external sources in source-based writing in l1 and l2. *Studies in Second Language Acquisition*, 41(3):555–582, 2019.

[7] Aaron D Likens, Laura K Allen, and Danielle S McNamara. Keystroke dynamics predict essay quality. In *CogSci*, 2017.

[8] Eva Lindgren and Kirk Sullivan. *Observing writing: Insights from keystroke logging and handwriting*, volume 38. Brill, 2019.

[9] Srdan Medimorec and Evan F Risko. Pauses in written composition: On the importance of where writers pause. *Reading and Writing*, 30:1267–1285, 2017.