Leveraging Natural Language Processing to Detect Gaming the System in Open-ended Questions in a Math Digital Learning Game

Jiayi Zhang[1], Shiyi Pang[1], J. M. Alexandra Andres[1], Ryan S. Baker[1], Elizabeth Cloude[1], Huy Anh Nguyen[2], Bruce M. McLaren[2]

[1]University of Pennsylvania
[2]Carnegie Mellon University
joycez@upenn.edu

Abstract
We present detectors that automatically identify when students game the system, a maladaptive learning strategy where students attempt to succeed by exploiting properties of a learning environment. In contrast to previous detectors that detected this behavior within students' interaction with learning activities, we detect within students' text-based responses to open-ended questions. With 5-fold student-level cross-validation, the model reached an average AUC ROC of 0.815, demonstrating a reliable method for detecting gaming in open-ended questions.

Keywords: NLP, gaming the system, automated detector

Introduction

Gaming the system (e.g., systematically guessing or abusing hints) describes students attempting to succeed in an interactive learning environment by taking advantage of the environment's properties (Baker et al., 2008). These disengaged behaviors have been observed across platforms and found to be negatively associated with learning and long-term outcomes (e.g., Cocea et al., 2009). Gaming detectors have been developed to study why students game (Li et al., 2022) and inform real-time intervention at scale (Xia et al., 2020). So far, gaming detectors have captured this behavior from interaction with learning activities or sequences of short response answers, but gaming can also occur in students' open responses. This past work has not looked at the semantic meaning of students' text inputs, solely looking for patterns in their responses.

In this study, we leveraged machine learning and natural language processing to detect gaming using textual responses. We collected ground truth data using text replay coding and applied a universal sentence encoder to vectorize the data. We trained a neural network to make predictions of gaming and evaluated the model with 5-fold student-level cross-validation.

Methods

Student log data were collected from *Decimal Point*, a single-player web game designed to motivate middle-school students to learn decimal concepts (McLaren et al., 2017). Students wander through a virtual amusement park and play a variety of mini-games that incorporate decimal challenges, such as sorting decimals. In the current version of the game, students are prompted with an open-ended, self-explanation question after problem solving in each mini-game (McLaren et al., 2022). In these questions, students are asked to reflect on how they solve the problem and explain their reasoning in text. To assure that students expend at least minimal effort, the response needs to contain at least four words with at least one of the words from a relevant list (including common misspellings) that would legitimately be found in a correct explanation. Students can make multiple attempts and only move to the next question once the response

meets these criteria. We collected the text-based responses submitted by 212 students and delineated them into clips, with each clip containing all the attempts (responses) a student made at answering a self-explanation question. We also collected pre-test and post-test measures.

Collecting ground truth. To build a machine learning model, we used text replay coding, where log data is delineated into human-readable clips (Baker et al., 2006). Human coders examine each clip and code the student's behaviors, establishing ground truth on the presence of gaming behaviors.

Based on the conceptualization of gaming defined in Baker et al. (2008), we developed a codebook to guide coding, which consisted of the following criteria: 1) the degree of semantic difference between responses, 2) cycling through multiple answers/modifications to their responses, or 3) conceptual or functional change between responses (e.g., identifying a concept versus suggesting an action) accompanied by the previous two criteria. Examples are given below:

Text Response Gaming Criteria	**Attempt 1**		**Attempt 2**	
Minor Semantic Difference	"I need to move it vertically"		"Move side to side"	
Cycling through Modifications	"It will be 7.1"		"It will be 7.2"	
	Attempt 1	**Attempt 2**	**Attempt 3**	**Attempt 4**
Conceptual or Functional Change	"It is 1.7"	"It is 1.9"	"By adding"	"By subtracting"

Using the codebook, the coder makes a holistic evaluation of whether gaming is present in each clip. Inter-rater reliability ($k = 0.8$) was established using a previous data set from the same platform. Once a clear consensus was reached within the research team, the assigned coder coded a total of 480 clips from 133 students. In these clips, only 6% of the clips (N=30) were labeled as gaming the system (approximately the same ratio as in past work on gaming), which is substantially imbalanced and may bias the predictor towards the majority category. To remedy the issue of imbalance and to maintain the originality in data, we oversampled the data in the training set, using the synthetic minority oversampling method (SMOTE) (Chawla et al., 2002) to reach 20% positive clips in the training set.

Building the model. We then utilized natural language processing techniques to create a feature space. Specifically, we concatenated every textual response in a clip together and then applied the Universal Sentence Encoder large v5 (Cer et al., 2018), generating a 512-dimensional vector for each entry.

Using the vector representation and the ground truth data, we trained a neural network with one-hidden layer to make predictions on gaming. We evaluated the model with 5-fold student-level cross-validation. We then computed the Area Under the Receiver Operating Characteristic Curve (AUC ROC) for each of the five testing folds and averaged them across folds. Additionally, we conducted hyperparameter tuning using Bayesian optimization on the training sets to see if this improved performance.

Results and implications:

Under cross-validation, the base model reached an average AUC ROC of 0.815 across folds, and the hyperparameter-tuned model reached an average AUC ROC of 0.821, reliable performance on held-out folds. After applying the base model to the full dataset (2554 clips across the 212 students), we found students' detected frequency of gaming was negatively correlated to pre-test (r = -.24, p<.001) and post-test (r = -.22, p =.002), but unlike previous research, gaming frequency was not correlated to normalized learning gains (r = 0.08, p =.274).

In contrast to previous gaming detectors that rely on interaction data, our model leveraged sentence embedding to detect gaming based on the semantic meaning derived from text-based responses. We plan to investigate whether our model can be applied to other open-response contexts where gaming the system occurs. In general, by detecting gaming the system in this additional context, we enrich understanding of how broadly this phenomenon occurs and enable learning technology to intervene in addressing gaming in a broader range of contexts.

Acknowledgements

We acknowledge support from NSF#DRL-2201798.

References

Baker, R.S.J.d., Corbett, A.T., Roll, I., Koedinger, K.R. (2008). Developing a Generalizable Detector of When Students Game the System. *User Modeling and User Adapted Interaction*, 18(3), 287-314.

Baker, R.S., Corbett, A.T., & Wagner, A.Z. (2006). Human classification of low-fidelity replays of student actions. *Proc. 8th International Conf. ITS*, 29-36.

Cer, D., Yang, Y., Kong, S.Y., Hua, N., Limtiaco, N., John, R.S., ... & Kurzweil, R. (2018). Universal sentence encoder.

Cocea, M., Hershkovitz, A., & Baker, R.S. (2009). The impact of off-task and gaming behaviors on learning: immediate or aggregate? *Proc. 14th International AIED*, 507-514.

Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Li, Y., Zou, X., Ma, Z., & Baker, R.S. (2022). A Multi-Pronged Redesign to Reduce Gaming the System. *Proc.* 23rd *International AIED,* 334-337.

McLaren, B.M., Adams, D.M., Mayer, R.E., & Forlizzi, J. (2017). A computer-based game that promotes mathematics learning more than a conventional approach. *International Journal of Game-Based Learning*, 7(1), 36-56.

McLaren, B.M., Nguyen, H.A., Richey, J.E., & Mogessie, M. (2022). Focused self-explanations lead to the best learning outcomes in a digital learning game. *Proc.16th ICLS, 1229-1232.*

Xia, M., Asano, Y., Williams, J.J., Qu, H., & Ma, X. (2020). Using information visualization to promote students' reflection on" gaming the system" in online learning. *Proc.7th ACM Learning@Scale*, 37-49.