Exploring the use of GPT-3 as a tool for evaluating text-based collaborative discourse

Authors: Tanner Phillips, Asmalina Saleh, Krista D. Glazewski, Cindy E. Hmelo-Silver
Indiana University
tanphill@iu.edu, asmsaleh@indiana.edu, glaze@indiana.edu, chmelosi@indiana.edu

Authors: Bradford Mott, James C. Lester
North Carolina State University
bwmott@ncsu.edu, lester@ncsu.edu

ABSTRACT: Natural language processing (NLP) models have previously been used to classify and summarize student collaborative actions in various online and computerized learning environments. However, due to limitations related to insufficient or inappropriate training data, these models are limited in their applications and impact. In this study, we explore how a new model, GPT-3, summarizes student chat in a computer-supported collaborative learning environment. With only a sentence explaining the context of the learning environment and two training examples, GPT-3 was able to effectively extract and summarize student conversations (properly attributing states such as frustration and confusion), reliably synthesize statements not present in the source text, and effectively ignore extraneous noise in the student chat. We discuss how this summarization could be used to support teachers understanding of student collaboration in computer supported collaborative learning environments.

Keywords: Natural Language Processing, Collaborative Learning, Discourse

1 INTRODUCTION

To support collaboration, instructors need to understand information about social interaction among students as they engage in computer-supported-collaborative learning, game-based learning, and other forms of instruction that focus on supporting group work. Researchers have used natural language processing (NLP) to analyze collaborative actions which are captured as speech or text (Blikstein & Worsley, 2016). However, most of these models require extensive training, and are mainly applied post-hoc, meaning they cannot be easily used for real-time orchestration. Even when sufficient previous data exists, these types of models tend to “fail non-gracefully” (Roschelle et al., 2020) when the context of the data changes slightly. Some researchers have used pre-trained NLP models to accomplish this task. However, the success of pre models has been limited because these models are often trained in a specific context, such as Wikipedia or Twitter data that doesn’t generalize well to student generated text (Phillips et al., 2021). A new model, GPT-3, has the potential to change this. GPT-3 was trained on a corpus of 410 billion tokens drawn from a common, largely indiscriminate crawl of the internet and is ten times large than any previous NLP model (Brown et al., 2020). This paper explores the potential of GPT-3 as a tool to summarize student chat in a computer-supported collaborative learning environment using only two training examples.

2 METHODS

The data analyzed in this study was collected from four 12-to-14-year-old students participating in a collaborative game-based learning environment designed to teach ecosystems concepts. In the learning environment, students are on a field trip to a fictional island in the Philippines and conduct investigations on why tilapia in the local fisheries are sick. They gather evidence, and then work together to determine the cause of the sickness. One of the major features of the game is a virtual whiteboard where students can drag-and-drop their collected evidence, organizing it by topic and potential explanations. Student discussion is supported by an in-game chat feature where students generate explanations, discuss evidence, and engage in group inquiry.

GPT-3 has been shown to be successful at NLP classification, summarization, and completion tasks with no or few training examples (Brown et al., 2020). To understand how it might be utilized in our game-based learning environment, GPT-3 was presented with the follow prompt:

> *Summarize the following conversations between four students, (Eagle520, Jeepney520, Sun520 and Turtle520) and their tutor (wizard520) who are playing an educational video game.*

We then primed GPT-3 by manually summarizing the first 10 lines of code in two chunks. This method could theoretically be implemented in practice, with teachers summarizing the first several lines of chat at the beginning of the class and allowing GPT-3 to continue during the implementation.

After being presented with these two training examples, the text of student chat from our data was split into groups of approximately 500 *tokens* to conform with the length limitations of GPT-3, or approximately 300-400 words. GPT-3 was asked to summarize these chunks. A total of 18 summaries were created over the 702 lines of chat.

3 RESULTS

Overall, GPT-3 was effective at summarizing chat conversations and was robust against previously unseen and unusual language. Additionally, GPT-3 could correctly attribute abstract concepts (e.g., spamming the chat, confusion, frustration) to specific students, even when the exact words of the summary did not appear in the chat. Because GPT-3 is text-bound, it does not have a robust understanding of context, and can make inaccurate inferences (see examples). This limited understanding of context means GPT-3 cannot always discern between content that is considered important in the specific context of a collaborative educational game and everyday conversation.

Example 1: *"Summary: Eagle520 is spamming the chat. Sun520 and Jeepney520 are upset with Eagle's behavior. Wizard520 asks the students to talk more respectfully as a team."* In this example, over half of the chat lines were the student's avatar of Eagle typing "REEEEE" into the chat. Both Jeepney and Sun express frustration with this, though Sun expresses this abstractly, saying: "can we kick this kid." The word "spam" appears within the text ("he kind of spam tho"), however, the phrase "spamming the chat" was generated by GPT-3.

Example 2: *"Summary: Wizard520 welcomes the students to the board. Wizard520 says they can get started by dragging and dropping notes from their notebook into the different columns. Sun and*

Jeepney are confused about how to do this." Despite not having been explicitly told about the format of the whiteboard, GPT-3 properly abstracts the concepts of notes dragged into columns on a board. Additionally, it correctly attributed an abstract state to Jeepney and Sun using a word not present in the text (confusion). Jeepney states: "We need help," while Sun states "I DONT SEE ANY COLLUMS."

Example 3: *"Summary: Wizard520 tells the students that they can't vote for their own notes. Eagle is having a mental breakdown. Sun thinks Turtle is a secret agent who is secretly not a turtle."* In this example, we see again how GPT-3's limited understanding of context means it is not capable of discerning between content-focused and social conversations. It nearly directly quotes eagle ("I WILL HAVE A MENTAL BREAKDOWN") and does not understand this is a hyperbolic expression of frustration. Additionally, it does not recognize the social nature of Sun and Turtle's conversation.

4 DISCUSSION

This paper explores the potential of GPT-3 as an aid in formative assessment of students in technology-enhanced learning environments. With only two training examples, GPT-3 was able to accurately summarize student chat in a collaborative game-based learning environment. It was able to accurately attribute abstract states to students, such as frustration and confusion, and behaviors such as "spamming the chat." Because of its limited understanding of the context of the student chat, GPT-3 does not discriminate between certain topics and does not always recognize hyperbolic statements. However, in the case of Example 3, this limitation could potentially prove invaluable in understanding the nature of collaboration. This is because it is possible that conversations that appear unproductive are instrumental in helping students regulate negative emotions (Author, 2011). Rather than labeling student actions along a spectrum of productive and unproductive behaviors, the model provides a simplistic description of the social situation. In doing so, the model allows the teacher to make inferences about the nature of student collaboration. The limitations of GPT-3 in summarizing the text in this study were relatively obvious and easily understood. If these limitations were properly expressed to teachers, this model could represent a substantial asset to teachers. Being able to view summaries of student conversation in real time would allow teachers to better allocate their attention to frustrated and confused students and gain a insight into students' progress in technology-enhanced learning activities. By placing the teacher as an intermediary between the model inference and pedagogical interventions, we could protect against the model failing non-gracefully (Roschelle et al., 2020). Overall, GPT-3 has the potential to markedly increase the accuracy and utility of NLP models in real-time analysis of educational data.

REFERENCES

Blikstein, P., & Worsley, M. (2016). Multimodal Learning Analytics and Education Data Mining: using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, *3*(2), 220–238. https://doi.org/10.18608/jla.2016.32.11

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *ArXiv*.

Phillips, T., Saleh, A., Glazewski, K. D., Hmelo-silver, C. E., Lee, S., Mott, B., & Lester, J. C. (2021). Comparing Natural Language Processing Methods for Text Classification of Small Educational Data. *Companion Proceedings 11th International Conference on Learning Analytics & Knowledge*.

Roschelle, J., Lester, J., & Fusco, J. (2020). *AI and the Future of Learning: Expert Panel Report Suggested Citation Acknowledgements* (Issue November). https://circls.org/reports/ai-report.