Received: 13 December 2021 | Accepted: 20 April 2022

DOI: 10.1111/bjet.13227

ORIGINAL ARTICLE

British Journal of Educational Technology BERA

Building socially responsible conversational agents using big data to support online learning: A case with Algebra Nation

Chenglu Li[1] | Wanli Xing[1] | Walter Leite[2]

[1]School of Teaching & Learning, College of Education, University of Florida, Gainesville, Florida, USA

[2]School of Human Development and Organizational Studies in Education, College of Education, University of Florida, Gainesville, Florida, USA

Correspondence
Wanli Xing, School of Teaching & Learning, College of Education, University of Florida, Gainesville, FL 32611, USA.
Email: wanli.xing@coe.ufl.edu

Funding information
Institute of Education Sciences, Grant/Award Number: R305C160004; University Florida AI Catalyst Award, Grant/Award Number: -P0195022; University of Florida Informatics Institute Seed Grant

Abstract

A discussion forum is a valuable tool to support student learning in online contexts. However, interactions in online discussion forums are sparse, leading to other issues such as low engagement and dropping out. Recent educational studies have examined the affordances of conversational agents (CA) powered by artificial intelligence (AI) to automatically support student participation in discussion forums. However, few studies have paid attention to the safety of CAs. This study aimed to address the safety challenges of CAs constructed with educational big data to support learning. Specifically, we proposed a safety-aware CA model, benchmarked with two state-of-the-art (SOTA) models, to support high school student learning in an online algebra learning platform. We applied automatic text analysis to evaluate the safety and socio-emotional support levels of CA-generated and human-generated texts. A large dataset was used to train and evaluate the CA models, which consisted of all discussion post-reply pairs (n = 2,097,139) by 71,918 online math learners from 2015 to 2021. Results show that while SOTA models can generate supportive texts, their safety is compromised. Meanwhile, our proposed model can effectively enhance the safety of generated texts while providing comparable support.

© 2022 British Educational Research Association.

KEYWORDS
discourse safety, online discussion forums, responsible conversational agents, social support

Practitioner notes

What is already known about this topic

- Online discussion forums have been plagued by a lack of interaction among students due to factors such as expectations to receive no response and perceptions of topic irrelevance which lead to low motivation to participate.
- AI-based conversational agents can automatically support students' interactions in online discussion forums at a large scale, and their generated responses can be human-like, contextually coherent and socio-emotionally supportive.
- Unsafe discourse exchanges between students and conversational agents can be dangerous as identity attacks, aggravation and bullying behaviours embedded in discourses can disrupt students' knowledge inquiry and negatively influence student motivation and engagement. However, few educational studies have paid attention to the safety of conversational agents.

What this paper adds

- This study proposes and synthesized strategies to build AI-based conversational agents that automatically support online discussions with safe and supportive discourses.
- This study reveals the relationship between discourse safety and social support, suggesting supportive discourses can also be unsafe.
- This study enriches the literature on educational conversational agents by synthesizing a conceptual framework on discourse safety and social support, and by proposing concrete algorithmic strategies to improve the safety of conversational agents.

Implications for practice and/or policy

- Researchers and practitioners can adopt strategies in this study such as generation control, open-sourced models and public API services to evaluate students' discourse safety for early intervention or modify existing conversational agents to be safety-aware.
- Practitioners can utilize the proposed conversational agent to automatically support students both safely and socio-emotionally at a large scale.
- Practitioners should be cautious when examining social support with automatic analysis, as not all supportive texts are safe. While unsafe texts can provide emotional support, it does not justify their appropriateness in a learning environment.

INTRODUCTION

Online discussion forums, an important pedagogical and social platform in online learning, have been shown in educational studies to support students' collaborative learning through perspective exchange (Almatrafi et al., 2018), negotiation (Gašević et al., 2019) and assimilation (Coman et al., 2020). Studies have revealed that the cognitive and socio-emotional support embedded in students' interactions in online forums may enhance their engagement and achievement (Moore et al., 2019; Salter & Conneely, 2015). However, online discussion forums have been plagued by a lack of interaction among students due to factors such as the expectation to receive no response and the perception of the irrelevance of topics, which lead to low motivation to participate (Chiu & Hew, 2018; Ezeah, 2014). Students' sparse interactions in online discussion forums can result in a vicious circle of low engagement where students can develop a sense of loneliness and suppression of sharing. Consequently, the low engagement level in discussion forums can prevent students from enjoying the benefits of the primary social setting (Tang et al., 2018) and lead to a high dropout rate (Ortega-Arranz et al., 2019).

To efficiently and effectively address students' low level of participation in online collaborative learning at a large scale, researchers have adopted learning design principles with learning analytics. Learning design focuses on the design and development of reusable learning activities through the creation and application of a repertoire of pedagogical tools (eg, taxonomy, frameworks) (Koedinger et al., 2013). Investigations include but are not limited to alignment evaluation (Zheng et al., 2020), participation level prediction (Er et al., 2019), learning pattern identification (Holmes et al., 2019), and real-time reports with teacher dashboards (Martinez-Maldonado, 2019). Besides using learning analytics methods, these studies constructed essential learning indicators through collaboration with teachers (eg, Er et al., 2019), referring to learning design frameworks (eg, quadratic assignment procedure in Holmes et al., 2019), or iteratively improving proposed systems with empirical examinations (eg, Martinez-Maldonado, 2019; Zheng et al., 2020). These learning indicators can provide automatic and actionable insights (eg, timing of learning activities, sequence of learning materials) to can assist teachers' individualized class orchestration and students' self-regulation.

Another promising approach in automatically supporting students' online collaboration are artificial intelligence (AI) based conversational agents (CA). CA and chatbot are two terms often used interchangeably (Syvänen & Valentini, 2020), which are defined as human-developed software powered by natural language processing techniques (NLP) to spontaneously respond to human languages (Wang et al., 2021). Most studies on AI-based CA in education have focused on responses' quality, engagingness, and learner experience (eg, Han & Lee, 2021; Li & Xing, 2021; Pereira et al., 2019). For example, Li and Xing (2021) built an AI-based CA to socio-emotionally support students in a massive open online course (MOOC). The results suggested that their AI-based CA could generate human-like responses and provide equal and sometimes greater socio-emotional support than that of humans. Pereira et al. (2019) created a chatbot capable of interacting with students in a MOOC by facilitating them with their assignments. Their results showed that most students (90%) reported having higher engagement to study with the chatbot-companion. In the study of Han and Lee (2021), the researchers built an English-based CA to answer students' frequently asked questions in a MOOC. The researchers showed that the effectiveness of the CA could be associated with various factors. For example, in their study, non-native English speakers reported significantly more perceived challenges in using the CA than native English speakers. Similarly, students located in Asia tended to report higher difficulties in using the CA as compared to those from other locations (eg, North America and Africa). However, few studies have paid attention to the safety of CAs. CA safety is operationalized as the evaluation of content appropriateness (eg, offensiveness) in generated texts (Dinan et al., 2021). Unsafe conversational exchanges between students and CAs can be dangerous as identity attacks and bullying behaviours embedded in discourses can disrupt students' knowledge inquiry and negatively influence students' motivation and engagement (Cruz, 2021; Pew Research Center, 2017; Trujillo et al., 2021).

This study aimed to address the safety challenges of CAs constructed with educational big data to support learning, which is the first step towards building socially responsible CAs. We intend to inspire future studies to use socially responsible CAs with learning design that incorporates individual differences as well as neurodiversity to support students' learning in online contexts. Specifically, we proposed a generic safety-aware CA model called SafeMathBot using strategies such as text generation style control to support high school students' learning in Algebra Nation, an online algebra learning platform. Details of safety strategies can be found in Methods. We compared the proposed model with two state-of-the-art (SOTA) CA models: BlenderBot (Facebook by Roller et al., 2021) and DialoGPT (Microsoft by Zhang et al., 2020). We applied automatic text analysis to evaluate the safety and socio-emotional support levels of CA-generated and original texts. In this study, original texts referred to posts or replies created by students and tutors in the investigated algebra learning platform. We used a large dataset to train and evaluate the CA models, which consisted of all discussion post-reply pairs (n = 2,097,139) by 71,918 online math learners from 2015 to 2021. The results show that while SOTA models can generate supportive texts, their safety is compromised. Meanwhile, our proposed model can effectively enhance the safety of generated texts while providing comparable support.

BACKGROUND

AI-based conversational agents

There are two distinct ways of constructing CAs. The first is a rule-based agent that requires manual engineering with classical NLP methods, and the other uses AI to generate responses with automatic data-driven inferences (Io & Lee, 2017). The former extracts keywords, intents, and emotions from students' input, producing responses with predefined templates. The latter often utilizes deep neural networks trained with *big data* to “learn” to respond to student input with human-like texts. Although both forms of CA can effectively support teaching and learning if constructed appropriately (eg, Grossman et al., 2019; Li & Xing, 2021), responses of rule-based CAs can suffer from input pattern coupling, topic limitations and wording repetitiveness. The limitations of rule-based CAs can be unwieldy in large-scale online learning courses and can potentially compromise students' learning experience (Jadhav & Thorat, 2020). Moreover, it is challenging for researchers to comprehensively consider different safety-related scenarios and manually construct predefined responses accordingly. Therefore, AI-based conversational agents seem to provide a more promising research direction and have attracted increasing attention.

In the context of AI-based CAs, the development of the transformer architecture has brought CAs to the next state-of-the-art (SOTA) level. Transformer (Vaswani et al., 2017) is a deep neural network architecture that solves many issues from prior deep learning models such as recurrent neural networks (RNN). For example, introducing the attention mechanism in transformer allows the model to inquire and understand which part of a sentence it should prioritize, which can help models to better capture contextual meanings of words and sentences (Devlin et al., 2019). Moreover, transformer is computationally efficient through parallelization, allowing researchers to construct larger models to better handle the high complexity of text data (Radford et al., 2019). Before transformer, researchers mainly used RNN to build CAs (eg, Indurthi et al., 2017; Tang et al., 2016). However, studies have shown that RNN-based solutions may not effectively retain information in long sentences (Wang et al., 2019). Furthermore, RNN tends to generate incoherent, hard-to-read and repetitive responses (Zhang et al., 2020), which can yield negative learning experiences. In contrast, studies with transformer-based CAs have shown impressive results in generating contextual and human-like texts (see the review of Zaib et al., 2020). Therefore, in this study, we have utilized the transformer to construct CAs to examine the affordances of SOTA models for text generation. Our proposed SafeMathBot and its benchmarks, BlenderBot and DialoGPT, are all transformer-based, the details of which are discussed in the Methods section.

Dimensions of support through the lens of social support theory

Social support theory enables researchers to examine how students share learning resources, assuming that the resource provider or receiver would benefit from the exchange (Shumaker & Brownell, 1984). Studies have found that students' learning gains (Goggins & Xing, 2016), relationship reciprocity and sustainability (Sconfienza et al., 2019), and learning engagement (Hsu et al., 2018) were positively associated with social support. Researchers often examined social support from three dimensions: informational, emotional and community support. These three dimensions of social support can be standalone or correlated with each other (Wellman & Wortley, 1990).

Students provide *informational support* by giving peers descriptive or substantive advice, suggestions, and insights (Wills, 1991). Researchers have extensively studied informational support in online contexts, showing that it could help reach desired learning results (Deetjen & Powell, 2016; Park et al., 2020; Xing et al., 2018). For example, Park et al. (2020) examined the role of informational support in an online healthcare community to support users' health resilience. Their results showed that constructive informational support could contribute to participants' goal-setting, which could influence their persistence in seeking health improvement in the community. Empathy, care, compassion, and reassurance are major forms of *emotional support* (Langford et al., 1997). Students' emotions have been shown to directly influence their level of participation and persistence in online learning contexts (eg, Hew, 2016; Xing et al., 2019). Moreover, students' perceived value of social presence and online learning was closely related to the emotional support that they received (Cleveland-Innes & Campbell, 2012). Finally, students provide *community support* through the invitation, recognition and maintenance of membership in a group, which can foster a sense of belonging, enhance involvement and improve motivation (Tomkin & Charlevoix, 2014; Wills, 1991; Zumbrunn et al., 2014). Student dropout rates can rise due to a sense of alienation brought on by a lack of community support (Almatrafi et al., 2018). This study examined the three dimensions of support provided by CA-generated texts using deep learning models via the perspective of social support theory.

Supportive discourses are not always safe

Whether an unsafe response is supportive or not can be subject to individual, contextual and cultural differences (Van De Poel, 2021). For example, for a biased person with a tendency towards racism or sexism, a biased response can be socio-emotionally supportive (see Table 1). However, for students who appreciate and accept the common values taught in schools, such unsafe responses can disengage and demotivate their learning (Cruz, 2021). Discourse safety is also a challenge for CAs, given the algorithmic bias that derives from human bias. Without specific handling, CAs can follow inappropriate human-generated texts and respond in an offensive way (eg, greetings, Q&A in Table 1). There is an estimation that 5–30% of online discourses are biased depending on the domain, which can explicitly and implicitly affect data-driven CAs (Curry & Rieser, 2018; Nobata et al., 2016).

Dinan et al. (2021) have proposed a framework to help researchers understand the three dimensions of discourse safety in the context of CAs, which can further shed light on why supportive discourses from CAs may not be helpful. In the framework, the authors proposed to examine CA safety in terms of its content appropriateness with instigator effect, yea-sayer effect and imposter effect, which were defined in their unique contexts. The *instigator effect* describes the initiation of offensive content by CAs. The *Yea-sayer* effect refers to the responses of CAs that explicitly or implicitly agree with offensive content from humans, which can fan the flame and lead to severe consequences such

TABLE 1 Examples of how social support and unsafety overlap

Human inputs	Human responses	Major social support	Discourse safety issue
- “bye have a nice day” - “may I post another problem”	- “shut up [Name]” - “[Name] can suck a cat”	–	Instigator effect
- “I have a big d” - “Math is so lame that i love it: D”	- “So am I :)” - “Wow. That's lame lol”	Emotional, community	Yea-sayer effect
- “[Name], can i please have some free karma points” - “When I multiply $7x^2$ and $2x^2$ what happens to the exponents?”	- “No, you must earn them, DUMMY” - “Damn we would add the exponents with x”	Informational	Instigator effect
- “nudes?” - “[Name] do u hate me now”	- “NO THATS GAY!!!!!!” - “omg just stop with your pointless drama ok hes my best friend okay just stop you did nothing wrong so stop acting and copying me thanks bye”	Community, emotional	Yea-sayer effect
What should I do if I got COVID?	Rest assured! COVID is not that serious	Emotional, informational	Imposter effect

Note: Quoted texts in the table were extracted from the dataset used in this study. These examples are infrequent and do not represent the opinions of the researchers or the studied platform. Unquoted texts were made-up examples due to a lack of appearance in the dataset and were out of the scope of examination in this study.

as violence and crimes. Finally, the *imposter effect* is the provision of incorrect or inappropriate advice in critical scenarios such as medical advice seeking, self-harming and emergency contexts. Generated content offensiveness of CA is a major benchmark in evaluating its three safety dimensions. Offensiveness includes but is not limited to sensitive, toxic, hateful and bullying content (Dinan et al., 2021), which highly overlaps with the six perspectives of evaluation (see Table 2) from Perspective API (2021a). Perspective API has been trusted by commercial products such as the New York Times and Disqus (Perspective API, 2021b). Numerous researchers have also examined Perspective API to perform automatic text safety detection and found the product robust and accurate (eg, Obadimu et al., 2019; Rieder & Skop, 2021). In this study, we will utilize Perspective API to evaluate CA content offensiveness to assist in the examination of the three safety dimensions.

There are studies using quantitative (eg, NLP) and qualitative (eg, human evaluations) methods to detect, evaluate and, to an extent, bypass these effects. For example, Mozafari et al. (2020) built a hate speech recognition model with deep neural networks to detect offensive content in social media, aiming to apply early interventions against undesirable behaviours and provide preventive insights to stakeholders. Lee et al. (2019) qualitatively analyzed how two open-sourced CAs responded to biased discourses. Their results showed that CAs could inappropriately agree with social biases such as racism and sexism. Finally, Xu et al. (2020) built a CA for healthcare, enhanced by a text classifier to detect users' purpose. If users intended to seek medical advice, their CA would generate texts with a templated response and direct users to expert-endorsed resources. Current attempts on CA safety focused on evaluating potential threats from user inputs, while few studies have proposed strategies to proactively enhance the safety of discourses generated by CAs. In this study, we aim to take the first step to fill this gap by proposing a method that can effectively enhance CA safety when generating discourses.

TABLE 2 Explanations and examples of safety evaluation with Perspective API

Safety perspectives	Descriptions	Examples
Toxicity	The extent of rudeness and disrespectfulness of a text that can cause a discussion to stop proceeding	"This is very easy. I can't believe someone does not know what a function is. LOL…"
Identity attack	The extent of negativity or hatred of a text that targets people's identity (eg, race, gender, sexual orientation)	"Ughh…girls!"
Threat	The extent of the intention to cause physical or mental harms against people	"Shut up or you will be in trouble!"
Profanity	The extent of using swearing, cursing, or other obscene language	"What the f**k does that mean?"
Insult	The extent of insulting and inflaming an individual or a group	"Please don't ask such a stupid question again"
Sexually explicit	The extent of references to sexual acts, body parts or other lewd content	"S*ck this!"

Note: Examples were real-world snippets generated by humans or CAs from this study's dataset.

Conceptual framework & research questions

Our conceptual framework can be found in Figure 1. Specifically, previous research shows that informational, emotional, and community support can overlap, while the three effects of CAs reside in mutually-exclusive contexts and are equally important. In the meantime, both safe and unsafe texts can be socio-emotionally (un)supportive. Safe and supportive texts can contribute to students' learning, while the opposite can negatively influence students cognitively and emotionally. Guided by the conceptual framework, we asked the following research questions:

1. To what extent can the proposed strategies enhance CA safety in the case of Algebra Nation? Details of proposed strategies can be found in the Methods section. Specifically, we evaluated the effectiveness of strategies developed in this study from the perspective of instigator and yea-sayer effects. The imposter effect was not examined as discussions of critical scenarios were extremely rare in our context.
2. To what extent can SafeMathBot provide social support to students in the context of Algebra Nation? We hypothesized that the formulated SafeMathBot could provide similar social support with its CA benchmarks and humans.
3. What is the relationship between discourse safety and social support in Algebra Nation? We hypothesized that there was no such relationship and correlation tests among the two would not yield any significance.

METHODS

Research context and dataset

We conducted this study within Algebra Nation (AN), a math learning platform used by 500,000 students across six states each year. We collected all the discussion posts and replies generated between 09/01/2015 and 09/01/2021 from the MySQL database of AN.

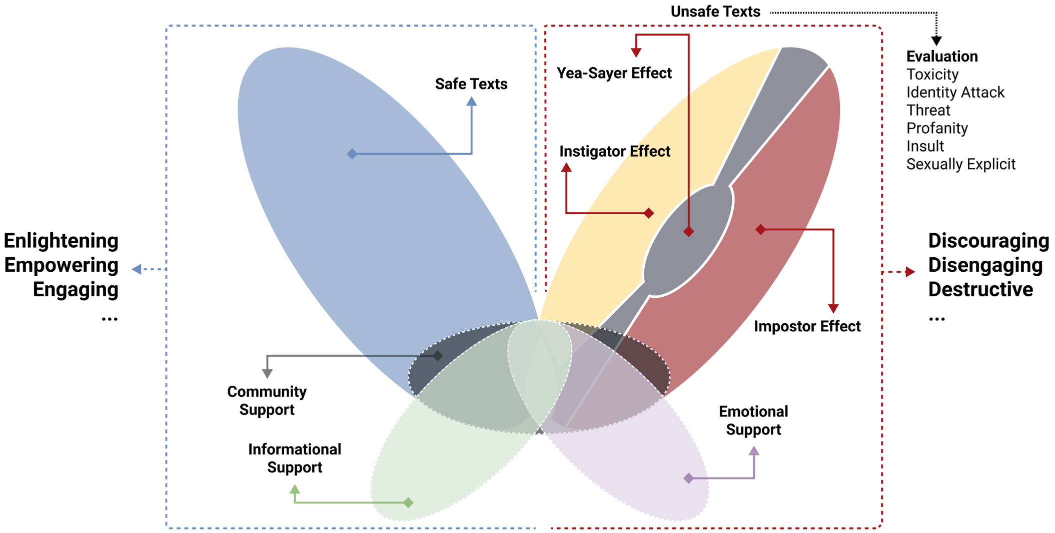


FIGURE 1 Conceptual framework to illustrate the relationship between discourse support and safety of CA. The shape or area of different components do not indicate their importance but are offered to provide a clear and aesthetic structure for interpreting the figure

After removing posts without replies, the dataset consisted of 2,097,139 post-reply pairs by 71,918 AN users. We operationalized a post as the initiation of an ongoing discussion and a reply as a response to a post. A reply can also be treated as a post if follow-up discussions were available based on this reply. There were 217,326 posts and 2,097,139 replies in the post-reply pairs, each post with an average of 9.65 replies (SD = 11.89). The discussion interactions on AN were from two parties, hired or volunteer tutors and students, where tutors were appointed to help answer students' questions. In the dataset, there were 69,465 (97%) students and 2453 (3%) tutors, where 205,989 (95%) posts and 1,529,281 (73%) replies were from students, with 11,337 (5%) posts and 567,858 (27%) replies from tutors. Most students saved their demographic information for registration ($n_{\text{female}} = 28,798$, $n_{\text{male}} = 33,216$, $n_{\text{unavailable}} = 7451$), while such information is not available for tutors. The post-reply pairs served as the training and evaluation samples for building conversational agents contextualized in math learning.

Safe conversational agents

Intuition

We developed the SafeMathBot by modifying generative pretrained transformer 2 (GPT-2, Radford et al., 2019), whose full model was reported to generate high-quality texts that could be dangerous to society if misused. Transformer models designed for language-related tasks are also called language models. Most language models utilize auto-regression to generate texts. Conceptually, auto-regression is the mechanism to calculate the probability of the next generated word based on previously generated words or contexts. We used this mechanism to make SafeMathBot safety-aware.

To illustrate, we extended the embedding layer of GPT-2 by allowing it to recognize two special tokens: [SAFE] and [UNSAFE], inspired by Xu et al. (2020). There are two embeddings in GPT-2; one encodes information for words and another for word positions. An embedding is a matrix of latent vectors that represent information such as a word's meaning and the role of its position with high-dimensional numeric values. The introduction of special tokens in SafeMathBot allowed us to adjust its word probability distributions for text generation. We associated non-offensive texts with the token [SAFE] and offensive ones with [UNSAFE]. To achieve this association, we have used ParlAI by Facebook (Miller et al., 2017) to classify whether a reply was safe or not automatically. An unsafe reply would yield offensiveness in the following perspectives: hate speech, personal attack, and profanity. The safety classifier was reported to achieve a prediction accuracy of 81.6 (Xu et al., 2020). To further examine its robustness, one author sampled replies ($n = 200$) predicted as safe or unsafe in our dataset, with 100 entries in each category. The other author rated the sampled texts' safety without knowing the prediction results. A good interrater reliability of 0.83 between the author and safety classifier was achieved and suggested the safety classifier was effective in our context. We then used the safety classifier to identify the safety of all the posts and replies in the dataset to let SafeMathBot be safety-aware. The original texts were suffixed with one of the special tokens of [SAFE] and [UNSAFE]. This process of adding extra information to data is called data augmentation (see Figure 2).

Safe discourse generation

After trained with texts enhanced by safety tokens, SafeMathBot, that can be “controlled” with these tokens to generate safe/unsafe texts on purpose. Figure 3 demonstrates the text generation process of SafeMathBot with safety control. In the example, we added a [SAFE] token at the beginning of a post. The post content was abbreviated as [POST], which could be “How is John doing with Algebra Nation?”. SafeMathBot then generated texts word by word, with “John loves math and he” being the current generation. To generate the next word, contexts and the current generation would first go through the embedding layers to extract information such as word meaning inferred from the training result. Then they would go through a stack of decoders. A decoder is a component in the transformer which consists of a self-attention layer and a feedforward neural network. The self-attention layer extracts the context information to allow models to know which part in a sentence tends to be more critical for response generation. At the same time, the feedforward neural network transforms data values to get rid of redundant information and retain the essential (Vaswani et al., 2017). Each decoder is the same except that the first decoder will take embedding values as input. In contrast, later decoders will take output values from the feedforward neural network of the previous decoder. In the example, the model generated the next word as “has”, given that this word has the highest probability among all the words known by the model.

Benchmarks

This study used two other CA models that have achieved SOTA performance in text generation, BlenderBot and DialoGPT, as benchmarks for SafeMathBot. As transformer-based CA models, both BlenderBot and DialoGPT shared a similar architecture with SafeMathBot, shown in Figure 3. However, their embeddings were not modified to accept special safety tokens and thus could not be controlled for safety. BlenderBot was innovative in its blending techniques that allowed the agent to learn different skill sets, such as demonstrating empathy and consistent personality (Roller et al., 2021). DialoGPT was based on GPT-2 and pretrained with 147 million conversational turns to generate meaningful and contextual texts (Zhang et al., 2020). Pretraining is another primary advantage of transformer-based models, where researchers or corporations train models with

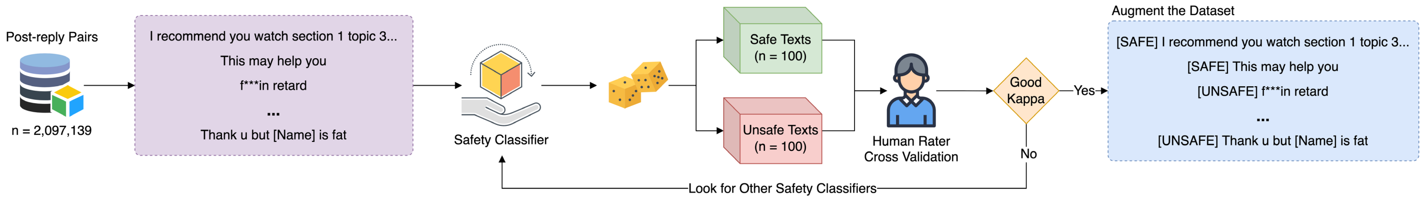


FIGURE 2 Data augmentation process to support safety-aware model learning

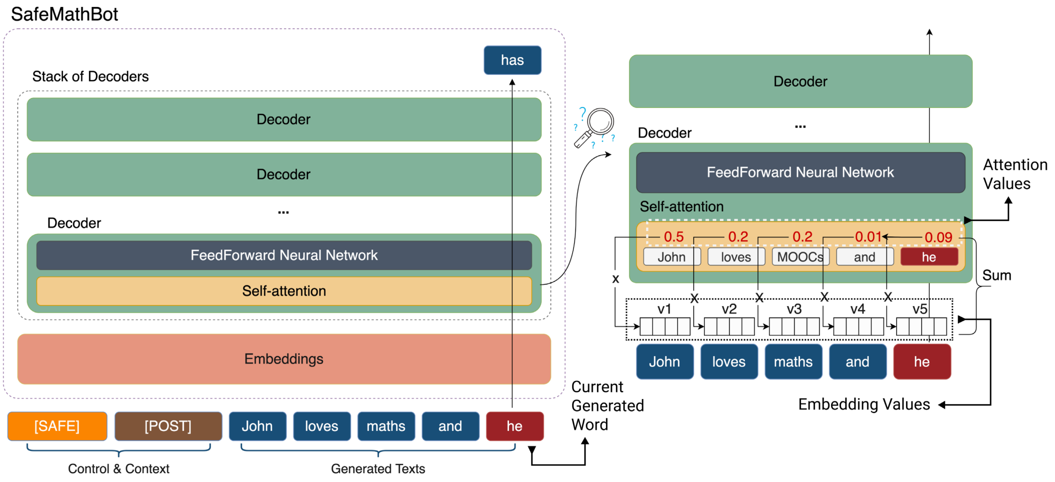


FIGURE 3 Demonstration of SafeMathBot's text generation process

a large amount of data using superior computing power. Model trainers then publish pretrained model weights for public use. Personalized datasets can then be used to fit pretrained models (called finetuning) by researchers. This process allowed custom-fit models to inherit impressive base performance on tasks essential for text generation such as natural language understanding (Fedus et al., 2018). GPT-2, BlenderBot, and DialoGPT offered different pretrained models, with larger sizes having better text generation ability. We conducted this study by using the largest pretrained model size available of all three models: GPT2/SafeMathBot (1.5 billion parameters), BlenderBot (3 billion parameters) and DialoGPT (762 million parameters).

Automatic text analysis of students' social support

Data processing & labeling

To understand students' social support expressed in their posts and replies, we have adopted advanced natural language processing (NLP) models to automatically detect such support in terms of informational, emotional and community dimensions. We first randomly sampled 1200 posts and replies from the entire dataset. Then 400 entries were randomly sampled from the random dataset. Two coders then independently coded the same 400 entries of the sampled data to ensure a high level of agreement could be reached. An entry of post or reply can be coded with multiple support. The degree of support ranges from 1 to 5, with 1 strongly disagreeing that a student has demonstrated the target support and 5 strongly agreeing with the provision of such support. The coders reached a high agreement with the qualitative coding measured with Cohen's kappa ($\kappa_{\text{informational}} = 0.89$, $\kappa_{\text{emotional}} = 0.87$, $\kappa_{\text{community}} = 0.88$). The two coders collaboratively solved disagreements and reached a consensus before conducting further coding. Finally, the two coders independently labelled 400 entries, respectively, which were merged as the training and evaluation data for NLP models.

Modelling

We have examined four NLP models to decide which model to select for further prediction. We constructed these models to perform separate regression tasks to predict the degree of informational, emotional, and community support. In this study, we have examined three traditional machine learning (ML) models that have been widely used in educational research and have shown promising results: Support Vector Machine (SVM), Decision Tree (DT) and Random Forest (RF) (eg, Hasan et al., 2020; Rizvi et al., 2019; Wiyono et al., 2020). We also examined a state-of-the-art deep learning model, Bidirectional Encoder Representations from Transformers (BERT, Delvin et al., 2019). The invention of BERT has greatly advanced the field of NLP. The learning mechanism and the adoption of the deep neural network of BERT have shown its superiority over traditional ML models in educational studies (Sung et al., 2021; Zou et al., 2021). This study used a base BERT model with 12-layers and 110 million parameters to perform the regression task.

Model training & evaluation

Model training and evaluation was conducted with the coded dataset (n = 1200), with 70% of data (n = 840) used for training and 30% used for evaluation (n = 360). We conducted 10-fold cross-validation with the three traditional ML models (SVM, DT and RF) to optimize performance on training and evaluation datasets. For BERT, 5-fold cross-validation was conducted by us, given that the large size of the model would make higher-order cross-validation challenging. To evaluate models, we used mean absolute error (MAE) and mean squared error (MSE) as metrics, which have been widely adopted in evaluating regression models. MAE is the average absolute distance between predictions and true values from the dataset, whereas MSE is the average squared distance.

Experiment setup

Figure 4 demonstrates how we set up and conducted the experiment.

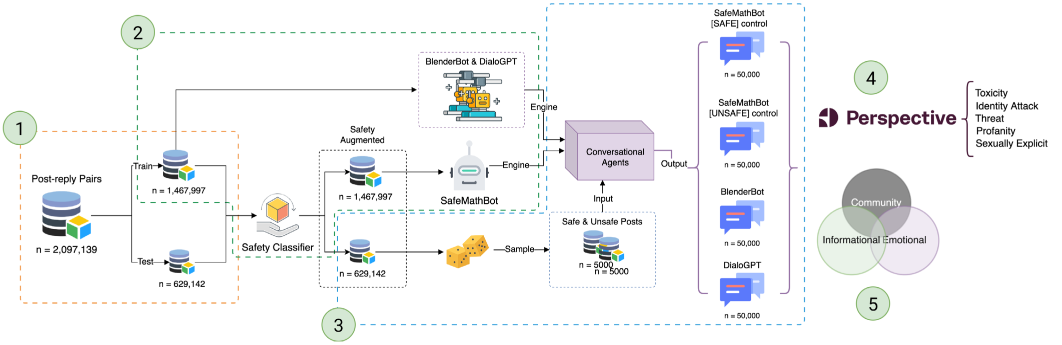


FIGURE 4 Illustration of the experimental process. Circled numbers with a green background in the figure correspond to the enumerated textual descriptions above

1. We first split the full dataset into training (70%, n = 1,467,997) and testing (30%, n = 629,142) sets. SafeMathBot, BlenderBot and DialoGPT were finetuned with the training set and evaluated with the testing set.
2. We used a powerful machine with 128 gigabytes of CPU RAM and eight NVIDIA RTX 2080 Ti GPUs to train CAs. Parallel GPU computation was needed as all three CAs had hundreds of millions of parameters, making CPU-only computation inefficient. Python packages, PyTorch (Paszke et al., 2019) and HuggingFace's Transformers (Wolf et al., 2020), were used to provide the computing infrastructure with GPUs. For BlenderBot and DialoGPT, models were trained with raw post-reply pairs, while we trained SafeMathBot with the data augmented with safety tags.
3. To examine instigator and yea-sayer effects of CAs, we randomly sampled 5000 safe and unsafe posts, respectively (n = 10,000), including their replies. The safe posts were used to examine whether CAs would initiate unsafe discourses, while the unsafe posts were used to understand how CAs would respond to offensive discourses. We then used the 10,000 posts as input sources for the three CAs to generate responses. For SafeMathBot, we used [SAFE] and [UNSAFE] tokens to understand the effectiveness of safety control. Each CA generated five responses for one post to better control for the random effects of CA generation (n = 200,000).
4. To evaluate CA-generated texts' safety, we used Google's Perspective API. As it is not the goal of this study to develop innovative safety evaluation models, but a CA that can generate safe texts, we used Perspective API instead of building our own evaluation engine. Specifically, we evaluated text safety from five perspectives: toxicity, identity attack, threat, profanity, insult and sexually explicit (see descriptions in Table 2). The safety classifier for SafeMathBot was not used in the evaluation. We were concerned that using the classifier would give SafeMathBot unwanted advantages since it was exposed to the classifier's predictions in training.
5. Finally, we used the text classifier described in the Automatic Text Analysis of Students' Social Support section to detect CAs' informational, emotional, and community support in text generation.

RESULTS

Descriptive statistics of dataset safety

Overall, a majority of posts (n = 214,608, 98.7%) and replies (n = 2,070,057, 98.5%) in the full dataset were predicted as safe, while there were a notable number of posts (n = 2718, 1.3%) and replies (n = 27,082, 1.5%) categorized as unsafe. Figures 5 and 6 illustrate the distributions of students' and tutors' safe and unsafe posts/replies, respectively. For example, Figure 5a shows that the proportion of unsafe posts by males was 0.86% (n = 1763), while that of female was 0.34% (n = 698). A higher proportion of unsafe replies can also be found in males (0.89%, n = 13,575) than females (0.47%, n = 7198) (see Figure 5b). The results indicates that male students created more unsafe posts and replies than females in this study, given that female students had more posts or replies than males, while creating less unsafe content. Considering there were more male (n = 33,216) than female (n = 28,798) students in the dataset, we calculated average unsafe posts and replies in terms of gender. On average, 0.02 unsafe posts were created by female students, while 0.05 unsafe post content was created by male students. A similar pattern was identified in students' replies, with female, on average, creating 0.25 unsafe replies and male creating 0.41 unsafe replies. This also suggests the female students' lower tendency to create unsafe content than males. However, it is important to note that the tendency to create unsafe content may be similar in

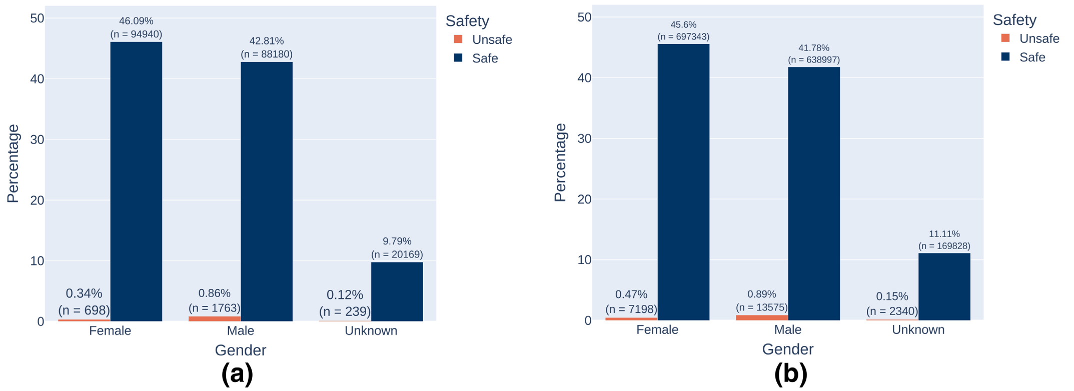


FIGURE 5 Students' discussion safety. Percentage values in each graph add up to 100%. In (a), each percentage value represents the proportion of safe/unsafe *posts* by gender based on all the *posts* of students. In (b), each percentage value represents the proportion of safe/unsafe *replies* by gender from all the *replies* of students

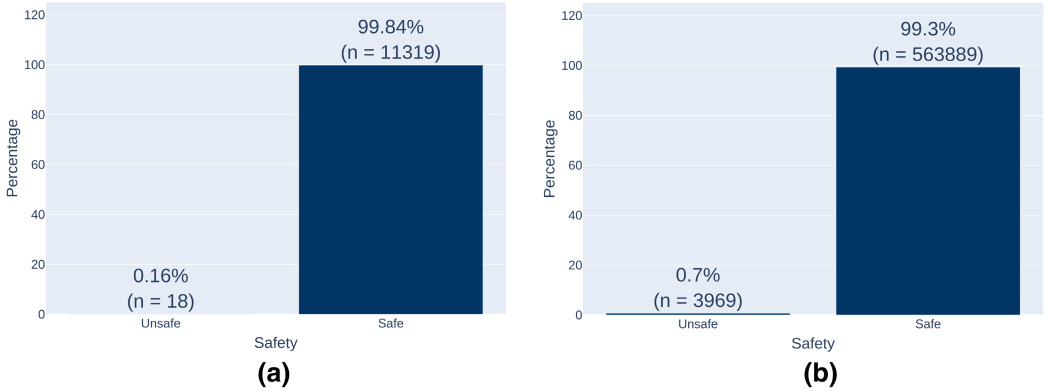


FIGURE 6 Tutors' discussion safety. Percentage values in each graph add up to 100%. In (a), each percentage value represents the proportion of safe/unsafe *posts* of all the *posts* from tutors. In (b), each percentage value represents the proportion of safe/unsafe *replies* of all the *replies* from tutors

posts and replies based on the similar proportions, although there were more unsafe replies than posts. Figure 6 shows that tutors tended to create less unsafe posts and replies than students, judging by the proportions. Table 3 provides examples of safe and unsafe posts/replies of students and tutors, where we can see the degree of tutors' discourses unsafety is less severe than that of students.

CA safety evaluation

Figures 7 and 8 show the histogram, kernel density estimate (KDE), and rug plot of KDE of CA-generated and original replies to safe or unsafe posts. A KDE line estimates the probability density function of a variable (eg, toxicity, identity-attack), which can be used to indicate how likely it is to see a value (eg, a probability of 0.7 for toxicity). A rug plot represents KDE of a variable in one dimension, which can assist with interpretations. The shared X-axis of histograms, KDE lines, and rug plots represents the probability of one safety issue being

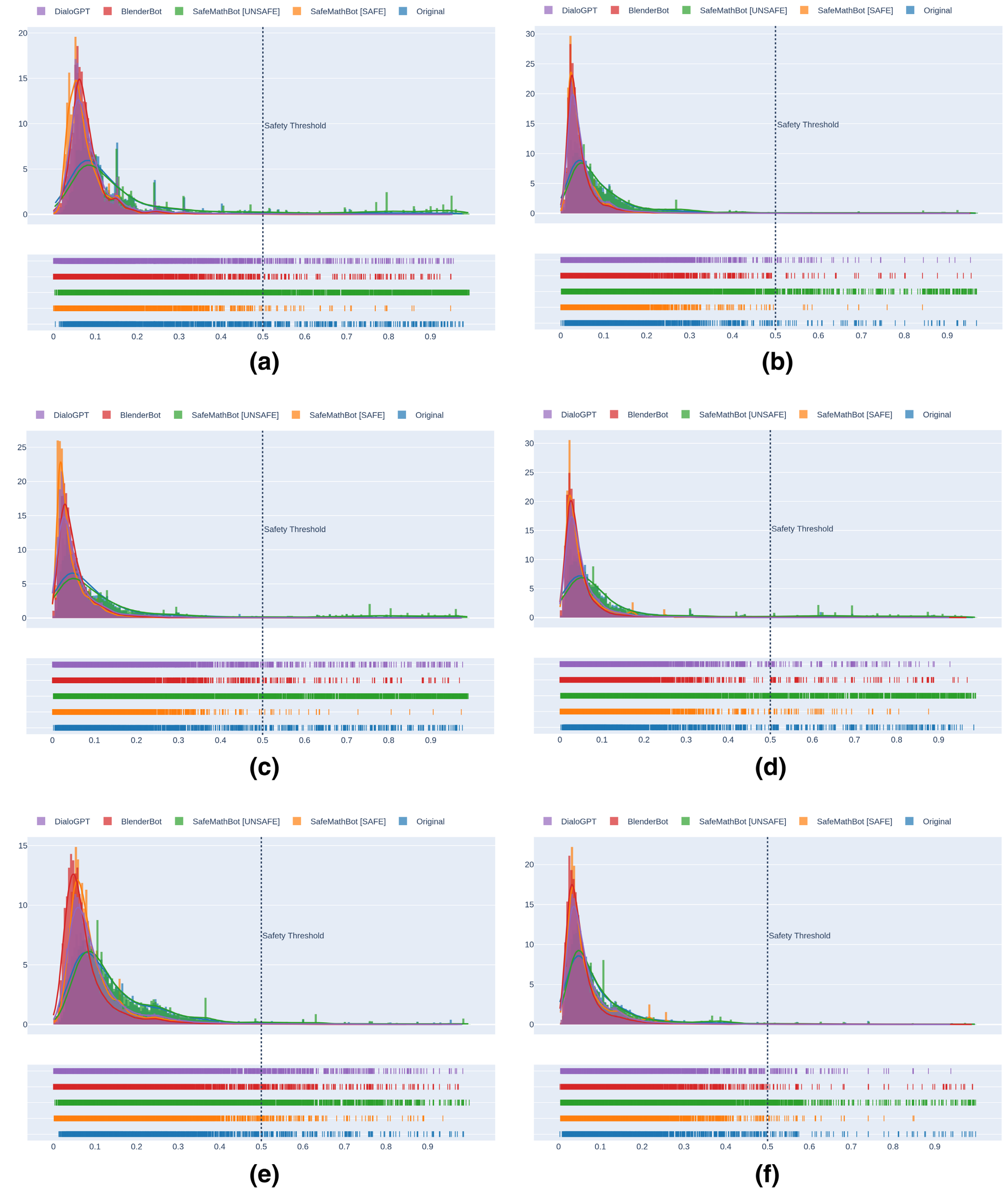
TABLE 3 Excerpts of students and tutors' (un)safe posts and replies

Role	Interaction type	Examples	
		Safe	**Unsafe**
Students	Posts	• "How do you multiply while using elimination? Can someone explain it for me?" • "Can an improper fraction be a irrational number?" • "Bye guys! Gotta go eat! I'll brb! :)"	• "Who else hates [Name]" • "[Name] you can Meet me in the streets. I'll pop the trunk on yo monkey self" • "Do any of you guys like carrot cake here? Personally I hate it and I think it should die"
	Replies	• "Yes [Name] I drew out the triangles" • "i can help too" • "i know how to do all the other problems but:) i dont know how to solve the problem if it has a varible in from of it"	• "u a thot" • "shut up" • "Your face is not Appropriate"
Tutors	Posts	• "Is it possible to get a workbook for Geometry that is translated into Spanish?" • "How do we get algebra nation to read the test questions to students?"	• "I need all of [Name]'s students to log off." • "How do i get access to my students password?"
	Replies	• "Hi [Name], do you have an Algebra question?" • "Recall if two lines are parallel, an intersecting line can form similar angles on the two lines"	• "Do you understand what you messed up on?" • "Please only speak about Algebra, not where you're from!"

Note: Example excerpts were extracted from the dataset to demonstrate post and reply safety of students and tutors.

true. A value of 0.5 for toxicity means the probability of a text being toxic is 50%. For example, the histogram and KDE in Figure 7a shows that most responses by CAs or humans had low predicted values of toxicity, with responses of SafeMathBot controlled with the [SAFE] token accumulated at a lower toxicity risk. Meanwhile, in the rug plot of Figure 7a, responses of humans and SafeMathBot controlled with the [UNSAFE] token were densely located above the safety threshold of 0.5, suggesting their responses had a higher risk of demonstrating toxicity. In contrast, SafeMathBot [SAFE] showed sparse rugs after the threshold, indicating a lower risk to generate toxic response. Both Figures 7 and 8 show that there was a trend for the following order in terms of safety: SafeMathBot [SAFE] > (safer than) DialoGPT > BlenderBot > Original Replies > SafeMathBot [UNSAFE].

To test this observation, we conducted a Pearson chi-squared test with Yates' continuity correction to examine whether the number of unsafe responses depended on who generated the response. We constructed a dependent variable to represent safety, which binarized six safety issues and converted them to one binary variable of safety. Specifically, the converted variable would have a 1 (safe) value if none of the six issues had a probability higher than 0.5. Otherwise, a 0 (unsafe) value would be assigned to the variable. Recall that we used CAs to generate five responses to each post. In the chi-squared test, we only kept the response with the highest generation probability among the five responses, with empty responses generated by CAs removed. The results of the chi-squared test showed that there was a significant association between safety and response creators ($\chi^2(4, 49{,}828) = 3744.4$, $p < 0.000$, Cramer's $V = 0.27$). The value of 0.27 for Cramer's V suggested a large effect size (Kim, 2017). Table 4 shows the crosstab between safety and response creators, which aligns with our observation from Figures 7 and 8 that SafeMathBot [SAFE] tended to generate more safe responses than the others.



(a) (b) (c) (d) (e) (f)

FIGURE 7 CA safety in response to *safe* posts to examine instigator effects using Perspective API. Histogram and KDE can be found in the upper region of a graph, while rug plot resides in the lower region. Less points residing on the right-hand side of the safety threshold suggest safer results. Original responses consisted of replies from both students and tutors in this study's dataset

CA social support evaluation

Table 5 shows the evaluation results of automatic text analysis for social support. We can see that BERT has achieved notably better performance than its benchmarks. We thus applied BERT to predict the level of support in CA- and original texts for the randomly sampled posts ($n = 10,000$). Table 6 provides examples of informational, emotional and community support

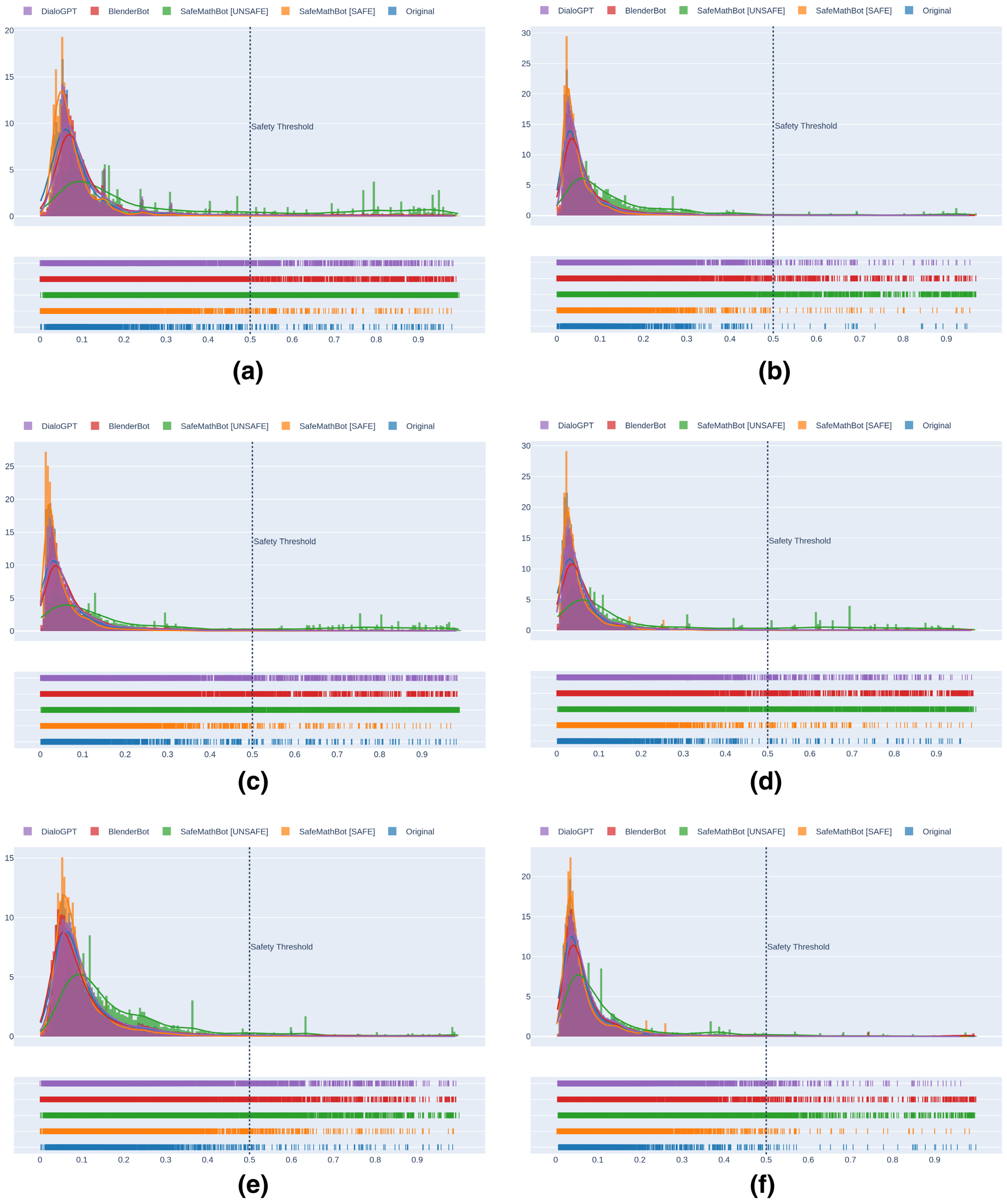


FIGURE 8 CA safety in response to *unsafe* posts to examine yea-sayer effects using Perspective API. Histogram and KDE can be found in the upper region of a graph, while rug plot resides in the lower region. Less points residing on the right-hand side of the safety threshold suggest safer results. Original responses consisted of replies from both students and tutors in this study's dataset

at varying levels. Figure 9 shows the distributions of social support of CAs and original replies. The plot outlines (eg, bell-shape curve) suggest the estimated probability of being in a specific support level, while the black dotted lines illustrate quartiles of data (25th, 50th and 75th percentiles). For example, we can observe that responses from both CAs and humans tend to provide informational support at a lower magnitude compared to community support.

TABLE 4 Crosstab of response safety and creators

	Safety		
Model	**Unsafe**	**Safe**	**Total**
BlenderBot	503	9497	10,000
	5%	95%	100%
DialoGPT	425	9522	9947
	4.3%	95.7%	100%
Original replies from students and tutors	781	9219	10,000
	7.8%	92.2%	100%
SafeMathBot [SAFE]	138	9801	9939
	1.4%	98.6%	100%
SafeMathBot [UNSAFE]	2267	7675	9942
	22.8%	77.2%	100%
Total	4114	45,714	49,828
	8.3%	91.7%	100%

Note: Each CA generated five responses to 10,000 posts ($n = 50,000$). Given there were 10,000 corresponding replies from students or tutors, only the one response with the highest generation confidence among five responses was kept in the Chi-squared test to retain a balanced sample size. CA response can be empty and responses without content were removed.

TABLE 5 Results of social support prediction models

Models	Metrics	Informational	Emotional	Community
BERT	**MAE**	**0.4963**	**0.6019**	**0.2863**
	MSE	**0.6621**	**0.6199**	**0.2651**
SVM	MAE	0.6652	0.6989	0.3013
	MSE	0.9859	0.7391	0.2664
DT	MAE	0.7440	0.7631	0.3759
	MSE	1.1225	0.9097	0.3554
RF	MAE	0.7492	0.7370	0.3557
	MSE	1.0225	0.7878	0.2912

Note: Lower values suggest better predictive accuracy. MAE (mean absolute error) estimates the average absolute distance between the estimator (eg, models) and estimated (eg, test data). MSE stands for mean squared error that measures the average of the squares of the difference between the estimator and estimated. The best-performed model is bolded.

In informational support, most responses fall within the magnitude range from 1 to 2, with a small portion of responses offering moderate (3) to strong (4 or 5) informational support.

We conducted a MANOVA test to examine whether there was a difference in average social support magnitudes among various response creators. The results showed that there was a significant difference ($F(12, 149{,}469) = 410.36$, $p < 0.001$ partial $\eta^2 = 0.032$). The small value of eta-squared suggested a trivial effect size and only 3.2% of the variance in social support magnitudes was explained by the type of response creators. Follow-up ANOVA tests showed that there were significant differences in informational ($F[4, 49{,}823] = 77.03$, $p < 0.001$, partial $\eta^2 = 0.006$), emotional ($F[4, 49{,}823] = 470.58$, $p < 0.001$, partial $\eta^2 = 0.036$), and community ($F[4, 49{,}823] = 796.58$, $p < 0.001$, partial $\eta^2 = 0.06$) support when inspecting the type of response creators. We conducted post-hoc tests with Fisher's least significant difference (LSD) tests to identify the order of average support magnitude from different response creators. The results can be found in Table 7, where we can see SafeMathBot

TABLE 6 Examples of social support at different levels

Support	Rating	Examples
Informational support	None (1)	• "I like you"
	Weak (2)	• "I guess not"
	Moderate (3)	• "yes but theres videos with the online teacher"
	Strong (4)	• "Yes. You can look at section 2 topic 3 if you have more questions. Good job"
	Very strong (5)	• "you are going to want to take the square root of both sides and cancel out the fraction in the denominator on the right side"
Emotional support	None (1)	• "[Name] that is not how you ask"
	Weak (2)	• "Sorry I got no clue"
	Moderate (3)	• "Hey, how are you doing today?? I am doing great!"
	Strong (4)	• "Thanks [Name]! You are amazing!"
	Very strong (5)	• "We have all been there before. Don't worry, you will understand it better… practice makes perfect!"
Community support	None (1)	• "shut up"
	Weak (2)	• "divide both sides by 9"
	Moderate (3)	• "Lmfaoooooo can i get it tho"
	Strong (4)	• "Greetings!!!"
	Very strong (5)	• "Bye, [Name]! Do you have anything to tell me about yourself?"

Note: Excerpts were extracted from the predicted dataset using BERT, which consisted of both CA-generated and original responses. Original responses consisted of replies from both students and tutors in this study's dataset.

controlled with the [SAFE] token can consistently provide comparable or greater support than humans and its benchmark CAs.

Relationship between safety and social support

We conducted point-biserial correlation tests between discourse safety and social support. Specifically, we correlated the dichotomized values (negative or positive) of six safety issues with informational, emotional, and community support. The dichotomous values were determined with safety issues' probabilities, with those greater than 0.5 coded as 1 (positive) and those smaller than or equal to 0.5 coded as 0 (negative). The results showed that all six safety issues had a weak-small negative relationship with informational and community support (see Table 8). This means that discourses predicted with safety issues tend to have lower informational and community support. It is worth noting that the correlation coefficients related to community support were all below 0.1, which suggested a negligible relationship (Ruscio, 2008). When using point-biserial correlation coefficients as the item discrimination index in psychometrics, a higher threshold of 0.2 is used to indicate marginal acceptance (Essen & Akpan, 2018). The risk of safety issues was positively correlated with emotional support, suggesting unsafe discourses tend to show higher emotional support.

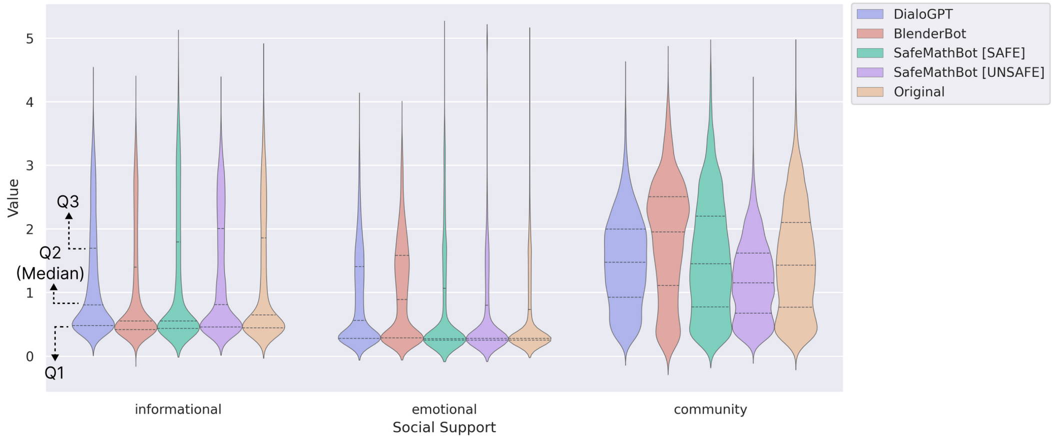


FIGURE 9 Distributions of social support of CA-generated and original replies. Original responses consisted of replies from both students and tutors in this study's dataset

TABLE 7 Placements of creators in terms of average support magnitude

Creators	Informational	Emotional	Community
Original replies from students and tutors	1	5	3
SafeMathBot [SAFE]	1	3	2
SafeMathBot [UNSAFE]	2	4	4
BlenderBot	3	1	1
DialoGPT	1	2	3

Note: Values in cells suggest the placement of support magnitude. For example, value 1 suggests the greatest mean magnitude of a certain type of support among all the creators.

DISCUSSION

AI-based conversational agents have attracted extensive attention in educational communities to provide automatic support for teaching and learning. Current educational studies on CAs mainly focused on their capability to generate responses of quality or users' experience with them (eg, Li & Xing, 2021; Han & Lee, 2021; Pereira et al., 2019). Another critical issue of CAs that lacks investigation in education is how to enhance their safety. In this study, we aimed to build socially responsible CAs for education with big data and artificial intelligence. Specifically, we have proposed an AI-based CA that utilized special tokens to learn to differentiate the boundary of safety. To the best of our knowledge, this study is the first of its kind to proactively address CA safety issues in education.

To reveal potential safety issues in online learning environments, we first conducted a descriptive analysis on this study's dataset. The results showed that there was a small portion of unsafe content generated by both students and tutors. There was a trend that male students generated more unsafe content than female students. This finding aligns with Bae's study (2021) examining teenagers' creation and exposure to cyberbullying content in online settings, where the researcher found that male students tended to create more offensive content than females. This may be explained by a relatively higher correlation between male students and violence before adulthood (Guo, 2016). Interestingly, Bae also found that students' cyberbullying behaviours were positively correlated with their exposure to unsafe content, which can imply the importance of creating socially responsible CAs in education.

TABLE 8 Point-biserial correlation tests between safety and social support

	Informational	Emotional	Community
Toxicity	**−0.15*****	**0.2*****	−0.02***
Identity attack	−0.07***	0.03***	−0.02***
Insult	**−0.14*****	**0.21*****	−0.02***
Profanity	**−0.13*****	**0.19*****	−0.03***
Threat	−0.03***	0.03***	−0.02***
Sexually explicit	−0.06***	0.03***	−0.01***

Note: Only correlation coefficients >0.1 are bolded, as values <0.1 are neglectable.
***$p < 0.001$.

The following sections discuss SafeMathBot's effectiveness in providing safe and socio-emotional support compared to safety-unaware CAs and humans. We also discuss the findings in examining the relationship between CA content safety and support level.

RQ1: Safety of SafeMathBot

Results have shown that SafeMathBot can effectively enhance the safety of its generations compared to other safety-unaware CAs. Specifically, Figures 7 and 8 show that SafeMathBot controlled with [SAFE] tends to respond more appropriately to safe and unsafe posts, showing a lower risk of instigator and yea-sayer effects. On the contrary, when controlling SafeMathBot with [UNSAFE], we find that the CA's generation is the least safe. The differing results between the two opposite controls of SafeMathBot suggest that we can effectively define the safety boundary with generation controls. Furthermore, BlenderBot and DialoGPT tend to generate more inappropriate responses, with BlenderBot generating more safe responses than DialoGPT. Our analysis with the chi-square test has further confirmed the safety enhancement of SafeMathBot, where SafeMathBot can generate almost six times fewer unsafe responses than humans and three times fewer than its benchmarks.

Previous educational studies have shown that toxic and offensive discourses could negatively influence students' learning experiences (Cruz, 2021; Pew Research Center, 2017; Trujillo et al., 2021). Even worse, exposure to unsafe content can stimulate depression and anger issues in students and can potentially convert victims to offenders due to imitation and pressure release (Lianos & McGrath, 2018). Furthermore, AI safety is important in establishing a trustworthy relationship among students, parents, teachers, and AI. A lack of trust in AI can impede its development and adoption in educational settings (Pedro et al., 2019). This study contributes to the current literature by proposing strategies to proactively enhance the safety of AI-based CAs. The proposed generation control method is generic and can be applied to other language models as long as their embedding layers can be extended. Practitioners can adopt strategies in this study such as generation control, open-sourced models, and public application programming interface (API) services to evaluate students' discourse safety for early intervention or to modify existing chatbots to become safety-aware.

RQ2: Social support of SafeMathBot

Results suggest that SafeMathBot can provide magnitudes of social support comparable to its benchmarks and to humans, which aligns with our hypothesis. Specifically, SafeMathBot showed the greatest average magnitude of informational support and achieved the second greatest average magnitude of community support (see Table 7). Although BlenderBot and DialoGPT had greater average emotional support than SafeMathBot, the small effect size suggests that the differences in emotional support among CAs are not practically meaningful. This finding aligns with our previous study which examined the use of a safety-unaware CA to support MOOC environments (Li & Xing, 2021). In the prior study, we have qualitatively analyzed the CA responses with human replies. A small-scale randomized experiment was also conducted with 4 participants to examine their perceived support from the CA and students. The results showed that the CA could generate texts, which are challenging to differentiate from those generated by humans, providing a similar level of support.

Numerous studies have shown that social support can effectively improve students' learning outcomes and enhance their learning retention (Goggins & Xing, 2016; Hsu et al., 2018; Sconfienza et al., 2019). This study contributes by proposing SafeMathBot to enhance its generation safety without sacrificing its capability to provide social support. Practitioners can utilize the proposed CA to automatically support students in online discussion forums at a large scale through open-domain conversations. Meanwhile, practitioners can build a separate widget or webpage with SafeMathBot in the form of a private chatting box. Finally, given the generic nature of the proposed strategy, practitioners can adopt the strategy to create safe CAs with other learning design theories to support students' online discussion activities. An example would be the integration of collaboration scripts (eg, Villasclaras-Fernández et al., 2009) to construct CAs with both rule-based and automatic text generations. Such CAs can provide a structured learning experience with templated learning activities and allow open conversations between students and CAs.

RQ3: Relationship between safety and social support

To our surprise, discourse safety is significantly associated with the magnitude of social support. However, the results are interpretable, especially when focusing on coefficients achieving meaningful sizes (eg, >0.1). Toxicity, insult and profanity are negatively correlated with informational support. This is understandable as these three aspects would usually contain negative attitudes with swearing words, which tend to digress and not provide the advice or suggestions needed in informational support (Infante & Wigley, 1986). On the other hand, these three safety aspects are positively correlated with emotional support. This finding resonates with Vanbrabant et al. (2012), where the researchers found that a mild to moderate level of verbal aggression is positively associated with people's social life. If we raise the probability threshold of unsafety to 0.8, the results suggest that none of the three safety aspects is meaningfully associated with emotional support ($\rho_{\text{toxic}} = 0.06$, $\rho_{\text{insult}} = 0.08$, $\rho_{\text{profanity}} = 0.03$). One possible explanation is that empathy is an important factor in emotional support, which can be safety agnostic. Whether a response is empathetic can depend on whether it helps soothe or alleviate others' negative emotions (Elliott et al., 2011). Therefore, using swearing words or being inappropriate does not prevent a response from being empathetic, thus emotionally supportive. An example from the dataset is one student posted, “the study expert is pretty hot”, and an unsafe reply predicted with moderate emotional support is “I know right? I wish I could have sl**t with her”. This explanation also applies to the finding that BlenderBot, specialized in providing empathy, shows an increased number of safety signs when responding to safe posts (see rug plots in Figure 7), and fewer when posts are unsafe (see Figure 8).

A general belief is that unsafe discourses are harmful from all perspectives (Vanbrabant et al., 2012). This study contributes to the literature by providing empirical results to show the relationship between discourse safety and social support from a different angle. Practitioners should be cautious when examining social support with automatic analysis, as not all supportive texts are safe. While unsafe texts can provide emotional support, it does not justify their appropriateness in a learning environment.

Limitations

One limitation of this study is the number of instances used to train the social support regressors. Thanks to the high transferability of BERT pre-trained with billions of texts, previous studies have shown a robust performance of BERT with small training sizes (eg, Sung et al., 2021). In the study of Sung et al., the researchers applied BERT for multi-label classification to understand students' scientific inquiry through their experimental reports ($n = 572$). Their results showed that BERT could achieve desirable predictive accuracy (AUC = 0.94) and greatly outperformed traditional machine learning models such as support vector machine by almost 5% accuracy. However, future research can benefit from more training samples to build automatic regressors with better generalizability.

Second, we treated tutors' and students' replies with the same weight for CA model training through the lens of social support. However, tutors' responses may be more knowledgeable and cognitively supportive as they received professional training. There is an opportunity to construct CAs with learning design based on tutor-generated content. Researchers can annotate the content of potentially high quality with concept map and learning activities (eg, following the conversational framework by Laurillard, 2013) to support students with individualized scaffolding and learning episodes using CAs as a terminal of communication.

Finally, although this study took the first step towards building a socially responsible CA for math learning with educational big data, we did not examine the CA in the actual learning context. It would be interesting and important to collect empirical data of students' experience of safety-aware and safety-unaware CAs to understand the effects of CA safety on students' learning. However, due to the sensitivity of CA safety, there might be concerns to conduct experimental studies in K-12 contexts. Future studies can consider higher educational settings or conduct randomized experiments with Amazon Mechanical Turk (MTurk). MTurk is a crowdsourcing platform with more than 500,000 potential participants and served as the platform for participant recruitment and payment in this study. Studies have suggested the reliability and robustness of the platform, which can be valuable for an exploratory study to extend educational community's knowledge on CA safety (eg, Paolacci et al., 2010).

CONCLUSION

A discussion forum is a valuable tool to support student learning in online contexts. However, interactions in online discussion forums are sparse, leading to other issues such as low engagement and dropout. This study proposed and synthesized strategies to build AI-based conversational agents that could automatically support online discussions with safe and supportive discourses. Furthermore, this study has shown that supportive texts are not necessarily safe in a learning environment, which can backfire on students' learning depending on context and individual differences. The findings imply that students can be facilitated with safe and socio-emotionally supportive discourses in large-scale online learning contexts. The ultimate goal of this study is to support students' interactions in online discussion forums by supporting them with CA-generated replies. To achieve the ultimate goal, apart from assessing CAs' ability to socio-emotionally support students, additional considerations are required when designing and evaluating them by incorporating theoretical frameworks such as Knowledge Building (Perkins, 2013) and Online Collaborative Learning Framework (Redmond & Lock, 2006) derived from Community of Inquiry to profoundly understand content quality of CA-generated texts and CA support dimensions (eg, teaching, cognitive, and social presence).

In the future, we plan to empirically investigate whether safety-aware CAs can effectively support students' learning. Following the Online Collaborative Learning Framework, we plan to design CA features (eg, hint messages based on students' learning inquiries, worked-example generation) to provide scaffolding and examine their affordances of teaching and cognitive presence. Moreover, understanding students' perceived safety of CAs can be equally important as understanding their algorithmic counterparts. Our previous study has suggested that algorithmic fairness might not significantly affect college students' perceived fairness towards an AI predictive system (Li & Xing, in press). Instead, factors such as individual differences (eg, majors, age) and transparency of system design were significant predictors of students' perceived fairness. To better design CAs for education, it can be important to reveal factors that influence students' perceived safety of CAs given that algorithmic fairness can be more nuanced (eg, challenging for layman to identify) than CA safety. Finally, most of the current approaches to assess discourse safety are fairness-unaware, having the potential to yield bias against students' identities. The bias can incorrectly associate predictions of lack of safety with trigger-words such as sexual orientation and racial descriptors (Borkan et al., 2019). The word *gay* can be toxic, neutral, or friendly depending on conversational contexts. However, there were findings that current safety classifiers tended to assign a high probability of toxicity to the word without considering its contextual meanings (Dixon et al., 2018). We plan to investigate and address the identity bias in discourse safety assessment using fair AI techniques (e.g., Li et al., 2021; Li et al., 2022).

ACKNOWLEDGEMENT

The research reported here was supported by the Institute of Education Sciences, US Department of Education, through Grant R305C160004 to the University of Florida, the University of Florida AI Catalyst Grant -P0195022, and the University of Florida Informatics Institute Seed Funding. The opinions expressed are those of the authors and do not represent the views of the University of Florida, Institute of Education Sciences, or those of the US Department of Education. We thank Dr. Kara Dawson from School of Teaching and Learning at the University of Florida for suggestions. We also thank Perspective API for allowing greater daily request quotas for this research.

ETHICS STATEMENT

The study was conducted in accordance with BERA Ethical Guidelines. No personal identifiers were reported in this study.

CONFLICT OF INTERESTS

There is no potential conflict of interest in the work.

DATA AVAILABILITY STATEMENT

The data set was collected with IRB approval from the authors' organization. These data will only be made available to other researchers if specific requests for amendments are made to the current approvals and will be considered on a case-by-case basis.

ORCID

Chenglu Li https://orcid.org/0000-0002-1782-0457
Wanli Xing https://orcid.org/0000-0002-1446-889X
Walter Leite https://orcid.org/0000-0001-7655-5668

REFERENCES

Almatrafi, O., Johri, A., & Rangwala, H. (2018). Needle in a haystack: Identifying learner posts that require urgent response in MOOC discussion forums. *Computers & Education*, *118*, 1–9. https://doi.org/10.1016/j.compedu.2017.11.002

Bae, S. M. (2021). The relationship between exposure to risky online content, cyber victimization, perception of cyberbullying, and cyberbullying offending in Korean adolescents. *Children and Youth Services Review*, *123*, 105946. https://doi.org/10.1016/j.childyouth.2021.105946

Borkan, D., Dixon, L., Sorensen, J., Thain, N., & Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of the 2019 World Wide Web Conference* (pp. 491–500).

Chiu, T. K., & Hew, T. K. (2018). Factors influencing peer learning and performance in MOOC asynchronous online discussion forum. *Australasian Journal of Educational Technology*, *34*(4), 16–28. https://doi.org/10.14742/ajet.3240

Cleveland-Innes, M., & Campbell, P. (2012). Emotional presence, learning, and the online learning environment. *The International Review of Research in Open and Distributed Learning*, *13*(4), 269–292.

Coman, C., Țîru, L. G., Meseșan-Schmitz, L., Stanciu, C., & Bularca, M. C. (2020). Online teaching and learning in higher education during the coronavirus pandemic: Students' perspective. *Sustainability*, *12*(24), 10367.

Cruz, S. (2021). *Cognitive and affective outcomes among targets and non-targets of racist hate speech in the college setting*. (Publication No. 2564152566) [Doctoral dissertation, Arizona State University]. ProQuest Dissertations & Theses Global.

Curry, A. C., & Rieser, V. (2018). # MeToo Alexa: How conversational systems respond to sexual harassment. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing* (pp. 7–14).

Deetjen, U., & Powell, J. A. (2016). Informational and emotional elements in online support groups: A Bayesian approach to large-scale content analysis. *Journal of the American Medical Informatics Association*, *23*(3), 508–513. https://doi.org/10.1093/jamia/ocv190

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186).

Dinan, E., Abercrombie, G., Bergman, A. S., Spruit, S., Hovy, D., Boureau, Y. L., & Rieser, V. (2021). *Anticipating safety issues in E2E conversational AI: Framework and tooling*. ArXiv.

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018, December). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 67–73).

Elliott, R., Bohart, A. C., Watson, J. C., & Greenberg, L. S. (2011). Empathy. *Psychotherapy*, *48*(1), 43–49.

Er, E., Gómez-Sánchez, E., Dimitriadis, Y., Bote-Lorenzo, M. L., Asensio-Pérez, J. I., & Álvarez-Álvarez, S. (2019). Aligning learning design and learning analytics through instructor involvement: A MOOC case study. *Interactive Learning Environments*, *27*(5–6), 685–698.

Essen, C., & Akpan, G. (2018). Analysis of difficulty and point-biserial correlation indices of 2014 Akwa Ibom State Mock Multiple Choices Mathematics Test. *International Journal of Education and Evaluation*, *4*(5), 1–11.

Ezeah, C. (2014). Analysis of factors affecting learner participation in asynchronous online discussion forum in higher education institutions. *Journal of Research & Method in Education*, *4*(5), 8–14.

Fedus, W., Goodfellow, I., & Dai, A. M. (2018). *MaskGAN: better text generation via filling in the_. ArXiv Preprint ArXiv:1801.07736*.

Gašević, D., Joksimović, S., Eagan, B. R., & Shaffer, D. W. (2019). SENS: Network analytics to combine social and cognitive perspectives of collaborative learning. *Computers in Human Behavior*, *92*, 562–577.

Goggins, S., & Xing, W. (2016). Building models explaining student participation behavior in asynchronous online discussion. *Computers & Education*, *94*, 241–251.

Grossman, J., Lin, Z., Sheng, H., Wei, J. T. Z., Williams, J. J., & Goel, S. (2019). *MathBot: Transforming online resources for learning math into conversational interactions. AAAI 2019 Story-Enabled Intelligence*.

Guo, S. (2016). A meta-analysis of the predictors of cyberbullying perpetration and victimization. *Psychology in the Schools*, *53*(4), 432–453. https://doi.org/10.1002/pits.21914

Han, S., & Lee, M. K. (2021). FAQ chatbot and inclusive learning in massive open online courses. *Computers & Education*, *179*, 104395. https://doi.org/10.1016/j.compedu.2021.104395

Hasan, R., Palaniappan, S., Mahmood, S., Abbas, A., Sarker, K. U., & Sattar, M. U. (2020). Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. *Applied Sciences*, *10*(11), 3894.

Hew, K. F. (2016). Promoting engagement in online courses: What strategies can we learn from three highly rated MOOCS. *British Journal of Educational Technology*, *47*(2), 320–341. https://doi.org/10.1111/bjet.12235

Holmes, W., Nguyen, Q., Zhang, J., Mavrikis, M., & Rienties, B. (2019). Learning analytics for learning design in online distance learning. *Distance Education*, *40*(3), 309–329.

Hsu, J.-Y., Chen, C.-C., & Ting, P.-F. (2018). Understanding MOOC continuance: An empirical examination of social support theory. *Interactive Learning Environments*, *26*(8), 1100–1118. https://doi.org/10.1080/10494820.2018.1446990

Indurthi, S. R., Raghu, D., Khapra, M. M., & Joshi, S. (2017, April). Generating natural language question-answer pairs from a knowledge graph using a RNN based question generation model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 376–385).

Infante, D. A., & Wigley, C. J., III. (1986). Verbal aggressiveness: An interpersonal model and measure. *Communications Monographs*, *53*(1), 61–69.

Io, H. N., & Lee, C. B. (2017, December). Chatbots and conversational agents: A bibliometric analysis. In *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 215–219). IEEE.

Jadhav, K. P., & Thorat, S. A. (2020). Towards designing conversational agent systems. In *Computing in Engineering and Technology* (pp. 533–542). Springer.

Kim, H. Y. (2017). Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. *Restorative Dentistry & Endodontics*, *42*(2), 152–155.

Koedinger, K. R., Booth, J. L., & Klahr, D. (2013). Instructional complexity and the science to constrain it. *Science*, *342*(6161), 935–937.

Langford, C. P. H., Bowsher, J., Maloney, J. P., & Lillis, P. P. (1997). Social support: A conceptual analysis. *Journal of Advanced Nursing*, *25*(1), 95–100. https://doi.org/10.1046/j.1365-2648.1997.1997025095.x

Laurillard, D. (2013). *Teaching as a design science: Building pedagogical patterns for learning and technology*. Routledge.

Lee, N., Madotto, A., & Fung, P. (2019). Exploring social bias in chatbots using stereotype knowledge. In *WNLP@ACL* (pp. 177–180).

Li, C., & Xing, W. (in press). Revealing factors influencing students' perceived fairness: A case with a predictive system for math learning. In *Proceedings of the Ninth ACM Conference on Learning@Scale (L@S '22)*. https://doi.org/10.1145/3491140.3528293

Li, C., & Xing, W. (2021). Natural language generation using deep learning to support MOOC learners. *International Journal of Artificial Intelligence in Education*, *31*(2), 186–214. https://doi.org/10.1007/s40593-020-00235-x

Li, C., Xing, W., & Leite, W. (2021). Yet another predictive model? Fair predictions of students' learning outcomes in an online math learning platform. In *Proceedings of the 11th International Learning Analytics and Knowledge Conference* (pp. 572–578). Association for Computing Machinery. https://doi.org/10.1145/3448139.3448200

Li, C., Xing, W., & Leite, W. L. (2022). Toward building a fair peer recommender to support help-seeking in online learning. *Distance Education*, *43*(1), 30–55. https://doi.org/10.1080/01587919.2021.2020619

Lianos, H., & McGrath, A. (2018). Can the general theory of crime and general strain theory explain cyberbullying perpetration? *Crime & Delinquency*, *64*(5), 674–700. https://doi.org/10.1177/0011128717714204

Martinez-Maldonado, R. (2019). A handheld classroom dashboard: Teachers' perspectives on the use of real-time collaborative learning analytics. *International Journal of Computer-Supported Collaborative Learning*, *14*(3), 383–411. https://doi.org/10.1007/s11412-019-09308-z

Miller, A., Feng, W., Batra, D., Bordes, A., Fisch, A., Lu, J., Parikh, D., & Weston, J. (2017, September). ParlAI: A dialog research software platform. In L. Specia, M. Post, & M. Paul (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 79–84). https://doi.org/10.18653/v1/D17-2014

Moore, R. L., Oliver, K. M., & Wang, C. (2019). Setting the pace: Examining cognitive processing in MOOC discussion forums with automatic text analysis. *Interactive Learning Environments*, *27*(5–6), 655–669. https://doi.org/10.1080/10494820.2019.1610453

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on BERT model. *PLoS ONE*, *15*(8), e0237861.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 145–153).

Obadimu, A., Mead, E., Hussain, M. N., & Agarwal, N. (2019). Identifying toxicity within youtube video comment. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation* (pp. 214–223). Springer.

Ortega-Arranz, A., Bote-Lorenzo, M. L., Asensio-Pérez, J. I., Martínez-Monés, A., Gómez-Sánchez, E., & Dimitriadis, Y. (2019). To reward and beyond: Analyzing the effect of reward-based strategies in a MOOC. *Computers & Education*, *142*, 103639. https://doi.org/10.1016/j.compedu.201

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, *5*(5), 411–419.

Park, I., Sarnikar, S., & Cho, J. (2020). Disentangling the effects of efficacy-facilitating informational support on health resilience in online health communities based on phrase-level text analysis. *Information & Management*, *57*(8), 103372.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*, 8026–8037.

Pedro, F., Subosa, M., Rivas, A., & Valverde, P. (2019). *Artificial intelligence in education: Challenges and opportunities for sustainable development* (Report No. ED-2019/WS/8). United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000366994

Pereira, J., Fernández-Raga, M., Osuna-Acedo, S., Roura-Redondo, M., Almazán-López, O., & Buldón-Olalla, A. (2019). Promoting learners' voice productions using chatbots as a tool for improving the learning process in a MOOC. *Technology, Knowledge and Learning*, *1–21*, 545–565. https://doi.org/10.1007/s10758-019-09414-9

Perkins, D. N. (2013). *Knowledge as design*. Routledge.

Perspective API. (2021a, November 15). *Using machine learning to reduce toxicity online*. https://www.perspectiveapi.com/

Perspective API. (2021b, November 15). *Case studies*. https://www.perspectiveapi.com/case-studies/

Pew Research Center. (2017). *Online harassment*. https://assets.pewresearch.org/wp-content/uploads/sites/14/2017/07/10151519/PI_2017.07.11_Online-Harassment_FINAL.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Redmond, P., & Lock, J. V. (2006). A flexible framework for online collaborative learning. *The Internet and Higher Education*, *9*(4), 267–276.

Rieder, B., & Skop, Y. (2021). The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API. *Big Data & Society*, *8*(2), 20539517211046181.

Rizvi, S., Rienties, B., & Khoja, S. A. (2019). The role of demographics in online learning; A decision tree based approach. *Computers & Education*, *137*, 32–47.

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E. M., Boureau, Y.-L., & Weston, J. (2021). Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 300–325).

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, *13*(1), 19–30.

Salter, N. P., & Conneely, M. R. (2015). Structured and unstructured discussion forums as tools for student engagement. *Computers in Human Behavior*, *46*, 18–25.

Sconfienza, C., Lindfors, P., Friedrich, A. L., & Sverke, M. (2019). Social support at work and mental distress: a three-wave study of normal, reversed, and reciprocal relationships. *Journal of Occupational Health*, *61*(1), 91–100.

Shumaker, S. A., & Brownell, A. (1984). Toward a theory of social support: Closing conceptual gaps. *Journal of Social Issues*, *40*(4), 11–36. https://doi.org/10.1111/j.1540-4560.1984.tb01105.x

Sung, S. H., Li, C., Chen, G., Huang, X., Xie, C., Massicotte, J., & Shen, J. (2021). How does augmented observation facilitate multimodal representational thinking? Applying deep learning to decode complex student construct. *Journal of Science Education and Technology*, *30*(2), 210–226.

Syvänen, S., & Valentini, C. (2020). Conversational agents in online organization–stakeholder interactions: A state-of-the-art analysis and implications for further research. *Journal of Communication Management*, *24*(4), 339–362.

Tang, H., Xing, W., & Pei, B. (2018). Exploring the temporal dimension of forum participation in MOOCs. *Distance Education*, *39*(3), 353–372.

Tang, S., Peterson, J. C., & Pardos, Z. A. (2016). Deep neural networks and how they apply to sequential education data. In *Proceedings of the Third ACM Conference on Learning@ Scale Conference* (pp. 321–324). ACM. https://doi.org/10.1145/2876034.2893444

Tomkin, J. H., & Charlevoix, D. (2014). Do professors matter? Using an a/b test to evaluate the impact of instructor involvement on MOOC student outcomes. In *Proceedings of the First ACM Conference on Learning@ Scale Conference* (pp. 71–78). ACM. https://doi.org/10.1145/2556325.2566245

Trujillo, M., Rosenblatt, S., de Anda Jáuregui, G., Moog, E., Samson, B. P. V., Hébert-Dufresne, L., & Roth, A. M. (2021). When the echo chamber shatters: Examining the use of community-specific language post-subreddit ban. In *Proceedings of the 5th Workshop on Online Abuse and Harms* (WOAH 2021) (pp. 164–178). https://doi.org/10.18653/v1/2021.woah-1.18

Van De Poel, I. (2021). Design for value change. *Ethics and Information Technology*, *23*(1), 27–31.

Vanbrabant, K., Kuppens, P., Braeken, J., Demaerschalk, E., Boeren, A., & Tuerlinckx, F. (2012). A relationship between verbal aggression and personal network size. *Social Networks*, *34*(2), 164–170.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, & R. Fergus (Eds.), *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6000–6010). ACM.

Villasclaras-Fernández, E. D., Isotani, S., Hayashi, Y., & Mizoguchi, R. (2009). Looking into collaborative learning: Design from macro-and micro-script perspectives. In *Proceedings of International Conference of Artificial Intelligence in Education* (pp. 231–238).

Wang, J., Hwang, G. H., & Chang, C. Y. (2021). Directions of the 100 most cited chatbot-related human behavior research: A review of academic publications. *Computers and Education: Artificial Intelligence*, *2*, 100023. https://doi.org/10.1016/j.caeai.2021.100023

Wang, Z., Ma, Y., Liu, Z., & Tang, J. (2019). *R-transformer: Recurrent neural network enhanced transformer. arXiv preprint arXiv:1907.05572*.

Wellman, B., & Wortley, S. (1990). Different strokes from different folks: Community ties and social support. *American Journal of Sociology*, *96*(3), 558–588.

Wills, T. A. (1991). Social support and interpersonal relationships. In M. S. Clark (Ed.), *Review of personality and social psychology, Vol. 12. Prosocial behavior* (pp. 265–289). Sage Publications, Inc.

Wiyono, S., Wibowo, D. S., Hidayatullah, M. F., & Dairoh, D. (2020). Comparative study of KNN, SVM and decision tree algorithm for student's performance prediction. *International Journal of Computing Science and Applied Mathematics*, *6*(2), 50–53.

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45).

Xing, W., Goggins, S., & Introne, J. (2018). Quantifying the effect of informational support on membership retention in online communities through large-scale data analytics. *Computers in Human Behavior*, *86*, 227–234. https://doi.org/10.1016/j.chb.2018.04.042

Xing, W., Tang, H., & Pei, B. (2019). Beyond positive and negative emotions: Looking into the role of achievement emotions in discussion forums of MOOCs. *The Internet and Higher Education*, *43*, 100690.

Xu, J., Ju, D., Li, M., Boureau, Y. L., Weston, J., & Dinan, E. (2020). *Recipes for safety in open-domain chatbots. arXiv preprint arXiv:2010.07079*.

Zaib, M., Sheng, Q. Z., & Emma Zhang, W. (2020). A short survey of pre-trained language models for conversational AI-a new age in NLP. In *Proceedings of the Australasian Computer Science Week Multiconference* (pp. 1–4).

Zhang, Y., Sun, S., Galley, M., Chen, Y. C., Brockett, C., Gao, X., Gao, J., Liu, J., & Dolan, W. B. (2020). DIALOGPT: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 270–278).

Zheng, L., Cui, P., & Zhang, X. (2020). Does collaborative learning design align with enactment? An innovative method of evaluating the alignment in the CSCL context. *International Journal of Computer-Supported Collaborative Learning*, *15*(2), 193–226.

Zou, W., Hu, X., Pan, Z., Li, C., Cai, Y., & Liu, M. (2021). Exploring the relationship between social presence and learners' prestige in MOOC discussion forums using automated content analysis and social network analysis. *Computers in Human Behavior*, *115*, 106582.

Zumbrunn, S., McKim, C., Buhs, E., & Hawley, L. R. (2014). Support, belonging, motivation, and engagement in the college classroom: A mixed method study. *Instructional Science*, *42*(5), 661–684. https://doi.org/10.1007/s11251-014-9310-0

How to cite this article: Li, C., Xing, W. & Leite, W. (2022). Building socially responsible conversational agents using big data to support online learning: A case with Algebra Nation. *British Journal of Educational Technology*, 00, 1–28. https://doi.org/10.1111/bjet.13227