International Journal of Artificial Intelligence in Education (2021) 31:186–214
https://doi.org/10.1007/s40593-020-00235-x

Natural Language Generation Using Deep Learning to Support MOOC Learners

Chenglu Li[1] · Wanli Xing[1]

Accepted: 15 December 2020/ Published online: 19 January 2021

Abstract

Among all the learning resources within MOOCs such as video lectures and homework, the discussion forum stood out as a valuable platform for students' learning through knowledge exchange. However, peer interactions on MOOC discussion forums are scarce. The lack of interactions among MOOC learners can yield negative effects on students' learning, causing low participation and high dropout rate. This research aims to examine the extent to which the deep-learning-based natural language generation (NLG) models can offer responses similar to human-generated responses to the learners in MOOC forums. Specifically, under the framework of social support theory, this study has examined the use of state-of-the-art deep learning models *recurrent neural network* (RNN) and *generative pretrained transformer 2* (GPT-2) to provide students with informational, emotional, and community support with NLG on discussion forums. We first trained an RNN and GPT-2 model with 13,850 entries of post-reply pairs. Quantitative evaluation on model performance was then conducted with word perplexity, readability, and coherence. The results showed that GPT-2 outperformed RNN on all measures. We then qualitatively compared the dimensions of support provided by humans and GPT-2, and the results suggested that the GPT-2 model can comparably provide emotional and community support to human learners with contextual replies. We further surveyed participants to find out if the collected data would align with our findings. The results showed GPT-2 model could provide supportive and contextual replies to a similar extent compared to humans.

This article belongs to the Topical Collection: *AI4MOOCs: Artificial Intelligence, sensoring, modeling and assessment for MOOCs. A step beyond*
Guest Editors: Filippo Sciarrone, Carla Limongelli, Olga C. Santos, and Marco Temperini

✉ Wanli Xing
wanli.xing@coe.ufl.edu

Chenglu Li
li.chenglu@ufl.edu

[1] Educational Technology, School of Teaching and Learning, University of Florida, Gainesville, FL 32601, USA

Keywords MOOCs · Natural language generation · Deep learning · Artificial intelligence · Discussion forums · Automatic support

Introduction

With rapid growth in popularity among learners and broad involvement with accredited organizations, massive open online courses or MOOCs have witnessed great success since their appearance in 2008 (Almatrafi et al. 2018; Babori et al. 2019; Xing et al. 2016). In December 2019, more than 110 million students enrolled in MOOCs from 13.5 thousand courses through mainstream MOOC platforms such as Coursera and edX (Shah 2019). Among all the learning resources within MOOCs such as video lectures and homework, the discussion forum stood out as a valuable platform for students' learning through knowledge exchange (Almatrafi et al. 2018; Wang et al. 2015; Xing et al. 2019). Studies have found peer support on various dimensions from discussion forums have positive effects on students' engagement and learning outcomes (Moore et al. 2019; Sunar et al. 2016). However, peer interactions on MOOC discussion forums are not prevalent (Chiu and Hew 2018). The lack of interactions among MOOC learners can then lead to a sense of isolation, suppressing learners' desire to share knowledge and concerns. In consequence, the low engagement level on discussion forums, to some extent, leads to MOOCs' long-standing issue of low participation and high dropout rate (Lee and Choi 2011; Ortega-Arranz et al. 2019; Xing and Du 2019).

As a result, extensive studies have been conducted to explore pedagogical innovations to spark MOOC learners' participation in the discussion forum. Conventionally, supporting MOOC learners on discussion forums much emphasizes the role of instructors and students. Researchers have investigated the pedagogical approach where instructors or teaching assistants (TAs) served as facilitators to connect with students and helped the course co-evolve with learners through the adaptive generation of resources and learning places (Kop et al. 2011; Masters 2011; Ruey 2010). Other studies focused on supporting MOOC learners from peers with the lens of connectivist. From the connectivist perspective, students are advocated to receive and provide support by showing autonomy, openness, connection with resources and people, and interactivity, typically in the form of discussion forum activities (Dubosson and Emad 2015; Mackness et al. 2013; Wise et al. 2017). Learning occurs through the evolvement of networks among students (Goldie 2016).

However, it is difficult, if not impossible, for instructors and learners to manually provide support to all participants on discussion forums given the vast quantity of learners on MOOCs (Almatrafi et al. 2018; Wise et al. 2017; Xing 2019). The recent trend has shifted to support MOOC learners on discussion forums with automatic text analysis methods. For example, Almatrafi et al. (2018) used machine learning models to automatically detect posts that needed urgent attention in order to allow instructors or TAs to assist more efficiently. Similarly, Wise et al. (2017) adopted machine learning algorithms to automatically classify whether posts bore meaningful content related to courses to help instructors and students find information aligned with their purposes more efficiently.

Among the automated methods of supporting the MOOC community, natural language generation (NLG) has shown potential in providing students with support in online learning settings (Caballé and Conesa 2019; Kumar and Rose 2010; Mittal et al. 2018). Though with multiple definitions of slight variation (Evans et al. 2002), in this study, NLG is defined as a subset of natural language processing (NLP) techniques that determine what is to be said meaningfully in response to language contexts in a particular language (McKeown 1982). Researchers have used NLG for a series of tasks such as conversation building, question answering, and reading comprehension (Benamara and Saint-Dizier 2004; Dale 2016; Du et al. 2017). In the study of Mittal et al. (2018), the researchers built a chatbot with NLG for MOOC learners and found the generated texts by the chatbot were able to offer informative answers to students' questions. MOOC learners were also found to benefit from NLG techniques in another study by Ferschke et al. (2015). In the study, Ferschke et al. found that the students who had access to a chatroom enhanced by the automated conversational agent were less likely to drop out as the course proceeded. Although studies have been done to explore the effects of using NLG to provide support for MOOC learners, to the best of our knowledge, most of the studies did not analyze the types of support NLG can offer to online learners compared with individuals. Furthermore, few studies have taken advantage of the rapid advancements in NLG with deep learning algorithms.

In this study, we explore the possibilities of using NLG with deep learning models to provide automatic support for students on MOOC discussion forums. The researchers have examined two state-of-the-art algorithms to generate replies to given posts, and they are *recurrent neural network* (RNN) and *generative pretrained transformer 2* (GPT-2). Studies have shown the potential of RNN to generate high-quality texts (Indurthi et al. 2017; Tang et al. 2016; Wang et al. 2018; Wen et al. 2015). In a study by Tang et al. (2016), RNN has shown a promising result through automatic word suggestions to provide students with writing support. GPT-2 is the state-of-the-art language model by OpenAI, the complete version of which was said by the authors to be too dangerous to release because of its capability to generate texts close to what can be generated by human writers (Radford et al. 2019). Therefore, we compared the performance of texts generation between the two promising NLG models to find an ideal candidate, with which we can analyze the possibility of automatically supporting students in MOOCs. Overall, this research aims to examine the extent to which deep-learning-based NLG can offer responses that are similar to human-generated responses to the learners in MOOC forums.

Background

Theoretical Foundations

Social support theory is the theoretical framework that helps researchers explore the resources exchanged by individuals with the perception by the provider or the receiver that such exchange would be beneficial to the receiver (Shumaker and Brownell 1984). Studies have shown a positive relationship between learning outcomes and social support (Hsu et al. 2018; Lin and Anol 2008; Tsai et al. 2017). Conventionally, research on social support was conducted within the physical environments to explore relationships between individuals and family members. However, with the increasing infiltration of digital devices and tools, much research has also begun to apply social support theory in digital environments (Tsai et al. 2017).

Social support theory can be distilled down to the following categories: informational support, emotional support, and community support (Cutrona and Russell 1990;

House 1981; Silverman 1999). Informational support yields the provision of advice, suggestions, and useful information to individuals (Wills 1991). Many studies have investigated informational support in online settings and have found informational support could help contribute to achieving intended outcomes (Deetjen and Powell 2016; Wu et al. 2019; Xing et al. 2018). For example, Xing et al. (2018) examined the effects of informational support on participants' continuing involvement with an online healthcare community. The findings suggested that informational support is positively associated with periphery users' commitment. Emotional support provides help in the form of empathy, concern, acceptance, and encouragement (Langford et al. 1997). In the context of MOOCs, studies found students' emotions directly related to students' engagement and persistence (Hew 2016; Jung and Lee 2018; Wen et al. 2014). In the meantime, emotional support dramatically contributes to how students perceive social presence and online learning (Cleveland-Innes and Campbell 2012). Finally, community support creates a sense of belonging by engaging in shared social activities (Wills 1991). Studies have suggested that the sense of community was positively correlated with students' engagement, motivation, and learning outcomes (Tomkin and Charlevoix 2014; Zumbrunn et al. 2014). A lack of belonging could lead to a sense of isolation that increased the dropout rate of students (Almatrafi et al. 2018; Lee and Choi 2011). Through the lens of social support theory, this study will examine the three types of support from texts generated with deep learning models.

Support of NLG for MOOC Learners

NLG techniques have been studied extensively in MOOC settings, mainly in the form of conversational agents (Caballé and Conesa 2019; Kumar and Rose 2010; Mittal et al. 2018). Conversational agents, also known as chatbots, engage users through natural language with the aid of algorithms (Shawar and Atwell 2007a). Much research has suggested conversational agents had the potential to support students in online learning (Pereira 2016; Pereira et al. 2019; Radziwill and Benton 2017; Shawar and Atwell 2007b). For example, Pereira et al. (2019) conducted a study with 77 students in a MOOC. The chatbot was designed to engage students by challenging them with questions and streamlining students' process of peer assignment review. The study found an improvement in students' motivation, with 90% of participants stating they would recommend the use of a chatbot in the future. In another study by Pereira (2016), the researcher designed a chatbot to help students with self-guided learning. Among the 23 participants, 89% of them thought the chatbot as a good companion for learning, and 72% suggested using the chatbot could help enhance engagement.

However, current research on the use of NLG to support learning has its limits. In a review of literature on chatbots, Io and Lee (2017) pointed out that most research limited the development of chatbots with classical NLP techniques, while few studies tapped in to the recent advancements of deep learning in NLG. The traditional NLG algorithms relied on the extraction of topics and generated predefined bodies of texts by humans (Demetriadis et al. 2018; Tegos et al. 2019). However, MOOCs are large-scale learning environments with a huge number of participants. Generating texts with limited topics and variations will be challenging to meet learners' needs. There is a need to empower conversational agents with responsiveness and creativity instead of solely relying on defined rules. The end-to-end nature of deep learning that requires few formal specifications for computers from human operators to achieve end goals (Goodfellow et al. 2016) makes deep learning an excellent candidate to support learning with NLG. The next section will present a review of deep learning for text generation.

NLG with Deep Learning Models

Deep learning often refers to deep neural networks. A deep neural network is "composed of multiple processing layers to learn representations of data with multiple levels of abstraction" (LeCun et al. 2015, p. 436). Compared with traditional machine learning approaches, deep learning shows a more promising future in that deep learning has proven its ability to process natural data in their raw form and requires little manual engineering to achieve desired results (Goodfellow et al. 2016; LeCun et al. 2015). Extant research has been conducted to explore NLG with deep learning, among which RNN stood out and showed promising results (Indurthi et al. 2017; Tang et al. 2016; Wang et al. 2018; Wen et al. 2015). For example, in the study of Indurthi et al. (2017), the researchers used the RNN to generate question-answer (QA) pairs given contexts (e.g., a body of texts describing a fact). The generated QA pairs were then compared with those generated with a non-deep-learning method. Their results suggested that RNN performed significantly better than the traditional method, and the generated QA pairs achieved a good quality overall. In another study exploring NLG with RNN, Wen et al. (2015) demonstrated that RNN could generate human-like texts on different topics.

In 2017, Vaswani et al. proposed the deep learning architecture *Transformer* for NLP, and continuous breakthroughs in NLP have been introduced since then (Devlin et al. 2018; Radford et al. 2019; Yang et al. 2019). The transformer architecture solves the computation bottleneck among prior deep learning models such as RNN and convolutional neural network (CNN) (Vaswani et al. 2017). Freedom to use vectorization computation in transformer architecture suggests better computational efficiency and allows researchers to experiment with models, including many more parameters. Recent models that achieved state-of-the-art performance on NLP are almost all based on the transformer architecture with hundreds of millions of parameters (Lan et al. 2019). GPT-2 is one of the transformer-based deep learning models, having achieved state-of-the-art performance in a series of NLP tasks such as question answering, reading comprehension, and summarization (Radford et al. 2019). The powerful language generation capability of GPT-2 has also been examined in studies across diverse domains (Budzianowski and Vulić 2019; Lee and Hsiang 2019; Zhang et al. 2020). For example, Lee and Hsiang (2019) adopted GPT-2 for augmented inventing. Specifically, the researchers trained a GPT-2 model to generate patent claims with 555,890 entries of sample claims. Their results showed that the patent claims generated by GPT-2 were coherent and reached a reasonable quality. In another example, Zhang et al. (2020) built a dialogue system with GPT-2. In the study, the researchers trained a GPT-2 and RNN model with conversational data from the *Reddit* public dataset to generate short responses to prompts. Through automatic and human evaluation, the researchers found that texts generated by GPT-2 could achieve a performance close to humans. In the comparison between GPT-2 and RNN, the results suggested GPT-2 outperformed RNN in terms of consistency with the dialogue context.

Furthermore, the architecture of transformer has been found to benefit researchers with pretrained models. Pretrained models are usually trained with a large amount of data by individuals or corporations with sufficient computation power. The parameters of pretrained models are then published for open use. Researchers can leverage the pretrained models to fit with custom datasets and inherit impressive base performance on tasks such as natural language inference and paraphrasing, which are essential for text generation (Fedus et al. 2018). Research has suggested that even models with limited training data can benefit from pretrained models (Lan et al. 2019). In the study of Budzianowski and Vulić (2019), the researchers found that the scarcity of training data for the NLG task can be mitigated by leveraging the pretrained model of GPT-2. The researchers evaluated results from GPT-2 automatically and manually and found that GPT-2 could generate high-quality texts even with limited training data.

From the literature reviewed, RNN and GPT-2 have shown great success in generating readable and coherent texts. However, limited research has been conducted to suggest the selection between RNN and GPT-2 when it comes to NLG. To the best of our knowledge, only one paper qualified with such a topic (Zhang et al. 2020) and no paper examined these two models in an educational setting. In order to select a more appropriate NLG model to better support students, in this study, we compared the performance of RNN and GPT-2 using automatic metrics.

Methods

Dataset and Research Context

The research context is a creative thinking course by a large U.S. public university in the northeastern area hosted on Coursera. The course included 7,066 students, with a large quantity of 42,307 posts and replies generated on the discussion forum during the course's first offering in the year 2013. The course lasted six weeks and intended to introduce its students to the creative thinking mindset so that students would be empowered to bring changes to their life and society through innovation. The course asked students to use the discussion forum to share their ideas on innovation and form groups to conduct group projects. The discussion forum also served as the place where students shared their group project results so as to receive peer feedback.

Data Preprocessing

In preparation for model training, the following data removal and cleaning procedures were conducted: (1) Removed posts without replies. The purpose of the modeling was to generate replies based on posts with deep learning, and posts without replies would not be helpful for model training. (2) Normalized textual content. Posts and replies were stored in the format of Hypertext Markup Language (HTML), and we used the Python package *BeautifulSoup* to clean up posts and replies so that HTML tags in the content would be stripped away. For example, a raw post such as "Hi<br />I think its great.<br />" would be converted to "Hi\nI think its great.\n", where <*br* /> means a line break in HTML and is replaced with the new line symbol *\n*. (3) Ensured 8-bit Unicode Transformation Format (UTF-8) encoding. UTF-8 is one of the most widely used text encoding formats that supports not only common characters but also non-character symbols such as emojis. To ensure all symbols in sentences could be accurately preserved, we used Python package *ftfy* to encode textual contents in UTF-8. (4) Masked web links. The deep learning models may learn the pattern of sharing web resources such as links when generating texts. However, the links generated by the models would either be taken from others' works or faked, which could be misleading and confusing. Therefore, we used regular expressions to target web links and replaced them with a special token *[LINK]* such that models can be informed to ignore the tokens during training.

After the data preprocessing, there were 13,850 entries of post-reply pairs left. Each entry of a post-reply pair contains one post and one corresponding reply. Posts would have duplicates since a post could have more than one reply (see Table 1).

Automatic Text Generation with NLG and Deep Learning

Overview

The deep-learning-based NLG models had three stages (see Fig. 1). In stage 1, data preprocessing filtered and transformed data to prepare it for model training by using the steps mentioned above. In stage 2, the processed dataset was split into a training, validation, and testing set with a proportion of 0.6 ($N = 8310$), 0.2 ($N = 2770$), and 0.2 ($N = 2770$), respectively. An RNN and GPT-2 model was trained with the training dataset; meaning model parameters were "learned" during this phase. Model architectures of RNN and GPT-2 are explained specifically in the next section. The validation set was used to estimate models' performance on an unseen dataset and evaluated if underfitting or overfitting occurred during model training. The testing dataset was used to provide contexts for the models to generate texts, the results of which were then evaluated quantitatively. In the last stage, models were evaluated based on word perplexity, readability, and cosine similarity with contexts (posts) and were compared with metrics of the original replies. Based on the quantitative metrics, one of the RNN

Table 1 Example of post-reply pairs

Posts	Replies
I'm interested in joining this study group. [Course Name] will be my first course at Coursera and looking forward to it.	I'm interested too
I'm interested in joining this study group. [Course Name] will be my first course at Coursera and looking forward to it.	I just sent my request to the group, thanks.
I'm re-reading "On Writing" by Stephen King. I recently finished "American Gods" by Neil Gaimon. I'll be focusing on the course's texts these next few weeks but I may be able to sneak in the next in line of the Mary Russell series by Laurie R. King - a gaslight mystery involving a retired Sherlock Holmes and his protege.	American Gods is a good book, But I think Neverwhere is Much better. Have you read it?

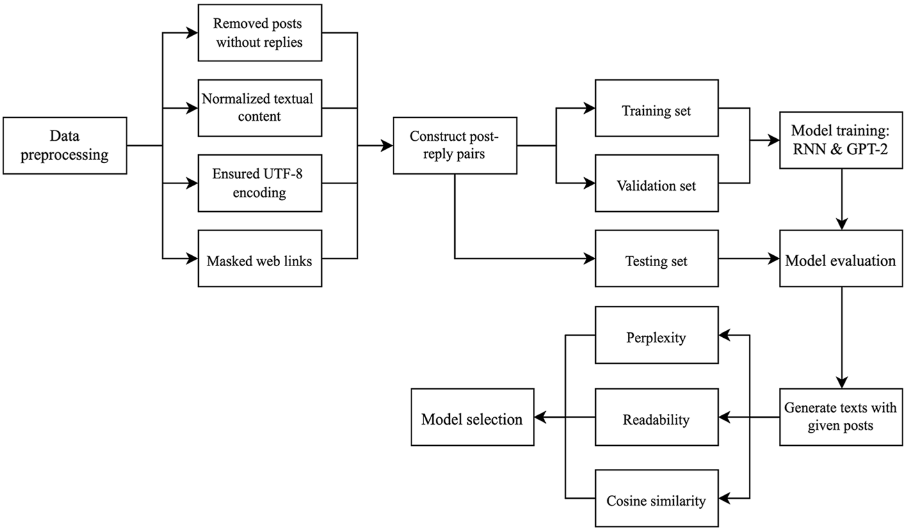

Fig. 1 Modeling procedures for NLG

and GPT-2 models was then selected for further qualitative analysis in understanding the types of support NLG can offer for MOOC learners.

NLG with Recurrent Neural Network (RNN)

RNN has been widely applied to explore its capability of text generation and has shown great potential in NLG tasks (Fedus et al. 2018; Mikolov et al. 2010; Potash et al. 2015; Sordoni et al. 2015). RNN is a variation of feedforward neural networks, which has three layers: an input layer, a hidden layer potentially consisted of multiple layers, and an output layer. RNN is different from a traditional feedforward neural network because RNN preserves data contexts through recurrent connections in the hidden layer, which significantly increases its performance on sequential data (Mikolov et al. 2010). The contexts in RNN, also known as hidden states, are retrieved through the current input (e.g., a word) and the information retained previously. The nature of relying on previous information to proceed defines the recurrence of RNN, which requires the model to process a series of data points one by one (see Fig. 2). As shown

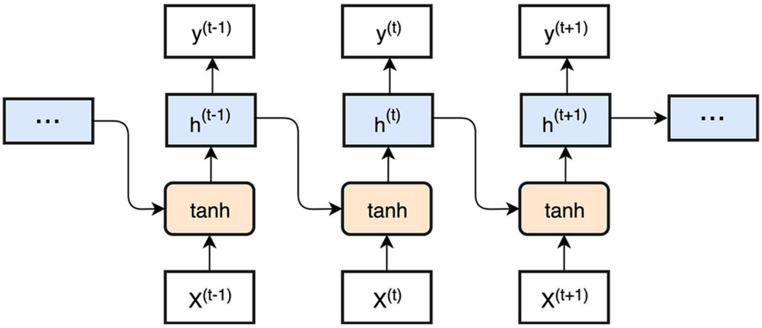

Fig. 2 Illustration of regular recurrent neural network for language generation (sequence-to-sequence)

in the figure, RNN takes a sequence of inputs X, with t denoted as the current timestep. To compute the hidden state h at time t, RNN needs to take the input of $x^{(t)}$and previous hidden state of $h^{(t-1)}$. The output y at time t is then generated based on $x^{(t)}$and $h^{(t)}$. The output is a probability distribution over the value of a given sequence.

However, training standard RNNs on problems that consist of long-term temporal dependencies (e.g., long sentences) can be challenging. A variation of RNN known as long short-term memory (LSTM) was found to be able to tackle the issues of long dependencies effectively (Graves 2013). LSTM improves regular RNN by introducing *memory cell*. The memory cell in LSTM can control when information should be preserved, when information should be output, and when information can be forgotten. Recent works using LSTM for language generation have suggested its great potential (Potash et al. 2015). Therefore, in this study, we used an LSTM-based RNN model for the NLG task.

NLG with Generative Pretrained Transformer 2 (GPT-2)

Generative pretrained transformer 2 (GPT-2) is a transformer-based deep learning model that achieves state-of-the-art performance on language generation (Radford et al. 2019). The GPT-2 model was pretrained with 40 gigabytes of texts by its authors, and four sizes of pretrained models have been released: (1) small with 12 layers and 117 million parameters; (2) medium with 24 layers and 345 million parameters; (3) large with 36 layers and 774 million parameters; and (4) full size with 48 layers and more than 1.5 billion parameters. Although larger GPT-2 models are presumed to have better performance, due to the limit of computational power, we chose the medium GPT-2 pretrained model with 24 layers and 345 million parameters in this research.

The transformer-based architecture of GPT-2 makes it different from RNN as it does not use recurrence in training to preserve contextual information (Vaswani et al. 2017). Due to this feature, when training with sentences, the transformer architecture does not need to process words one by one. The training process is thus much more efficient. However, many transformer-based models, including GPT-2, use the idea of auto-regression when generating texts, which resembles the mechanism of RNN. Auto-regression is the mechanism that when getting generated, words are output sequentially, and the newly generated word will be concatenated with previously generated words to form a new sequence. This new sequence will then serve as the input to generate the next word (Radford et al. 2019). Since this study addresses NLG, this section focuses on the text generation instead of the training mechanism of GPT-2.

Figure 3 demonstrates the architecture of GPT-2 for text generation, the values in which were made up for illustration purposes. In the left portion of the graph, an incomplete sentence, "John loves MOOCs and he", has been generated, and the model is predicting the next word after "he". Each generated word will first go through the embeddings layer, which will transform the words into numeric vectors. The embeddings layer contains two types of embeddings, a word embedding and a positional embedding. The two embeddings are matrices calculated during training and the model queries through a word or a word's position in the sequence to get associated vector values. The word embedding has 50,257 rows, with each row being a unique word from the vocabulary of the pretrained GPT-2 model. The length of columns varies depending on the size of the pretrained GPT-2 model. For example, the medium version has 1024

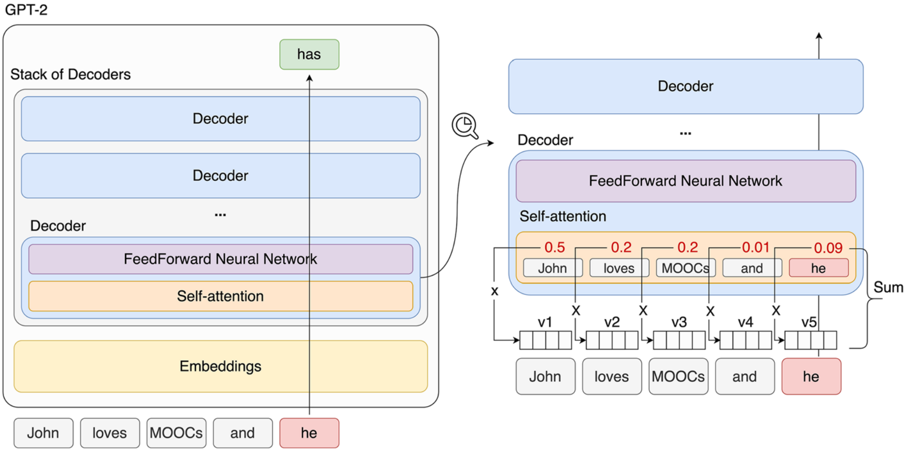

Fig. 3 Illustration of the GPT-2 model for language generation

columns, meaning a word will be transformed into a vector of length 1024. Words that cannot be found in the matrix will be using the vector of the special word “Unknown”. The positional embedding works similarly except that the rows are position numbers. For example, the word “he” in the figure is in position 5, so its value will be extracted from row 5 in the positional embedding. For each word, the embeddings layer will sum up the two vectors from the two embeddings and pass the result to a stack of decoders.

Each decoder in the stack has the same architecture while having different weights, which are calculated during training. Each decoder has two components, a self-attention layer and a feedforward neural network. The right part of Fig. 3 shows the mechanism of the self-attention layer. Conceptually, the self-attention layer is to get the contextual information related to the currently processed word. Specifically, each word will be assigned an attention score, with greater scores being more important to the currently processed word. Furthermore, each word has a vector value. For the first decoder (bottom one) in the stack, the vector values of words are the outputs from the embeddings layer. For other decoders, the vector values will be the outputs from the feedforward neural network in the previous decoder. The self-attention will sum up each word’s vector values multiplied by their attention scores, which will be fed into the feedforward neural network. The mechanism of self-attention is said to retain each word’s contextual information in a vector representation (Vaswani et al. 2017). The output from the self-attention layer is then fed into a feedforward neural network for feature extraction and dimension reduction. In the example of Fig. 3, “he” refers to “John”, so the expectation is the word “John” has the highest attention score. Since “he” is the currently processed word, its attention score tends to be small. After repeating the computation in each decoder orderly, the model will output a probability distribution of each word in its vocabulary. In our example, the word “has” achieved the highest probability and is thus predicted as the next word.

Evaluation

After training, models of RNN and GPT-2 were evaluated on the testing dataset with three metrics to aid researchers with model selection. We first generated replies with

RNN and GPT-2 by using the 2770 posts in the testing dataset. The posts in the testing set were used as “prompts” of models, meaning the RNN and GPT-2 models would generate replies corresponding to those prompts. Then word perplexity, Flesch-Kincaid (F-K) grade level, and cosine similarity were computed for evaluation. The F-K grade level and cosine similarity of the original replies were also computed to serve as the benchmarks for evaluation. Since word perplexity is a metric specifically for NLG models, the original replies thus do not have such measure.

Word Perplexity Word perplexity is a well-established metric for evaluating model performance in NLG (Bengio et al. 2003), which describes how surprised the model is when encountering the predicted next word (Evermann et al. 2016). For example, a word perplexity of 30 indicates the model proposes 30 words to choose from when generating the next word. In practice, lower word perplexity is preferred as it suggests a model is more confident in knowing what the next word should be (Serban et al. 2016). Word perplexity is defined as (Pietquin and Hastie 2013)

$$Perplexity = e^{\sum_{i=1}^{N}\frac{1}{N}\log(p_m(x_i))}$$

, where $p_m(x_i)$ is the probability of the word x_i given the model, and N is the number of possible words. However, since logarithm was computed with a base of the *Euler number* (natural log) in the models, the exponential function was used instead of a base of 2 in the formula.

Flesch-Kincaid (F-K) Grade Level for Readability Initially designed for the U.S. Navy to test readability of English texts (Kincaid et al. 1975), Flesch-Kincaid (F-K) grade level has been widely used in NLP as well as educational settings (Dufty et al. 2006; Feng et al. 2010; Xu et al. 2016). The F-K grade level metric is defined as

$$F{-}K\ Grade\ Level = 0.39\cdot\frac{total\ words}{total\ sentences} + 11.8\cdot\frac{total\ syllables}{total\ words}{-}15.59$$

The F-K grade level is an approximation to the grade level required to understand a given text. In this study, we used the Python package *textstat* to compute F-K grade levels of generated texts from the testing dataset.

Cosine Similarity for Semantic Coherence While word perplexity can indicate how well an NLG model performs and F-K grade level can suggest the readability of generated texts, there is a need to automate the evaluation of how coherent generated texts are with original posts. Because off-topic replies are not likely to be supportive and can be confusing for students. Traditionally, manual works will be applied in the evaluation of generated texts’ semantic coherence. However, with the advancements of NLP, recent studies have suggested that using word embeddings to compute cosine similarity can well explain generated texts’ coherence with conversational contexts (Clark et al. 2019; Fang et al. 2016; Vakulenko et al. 2018).

Word embedding is a technique that represents words with vectors such that textual words can have numeric representations that are processable by machine learning algorithms. Furthermore, word embedding retains word meaning and semantic context so that similar words would have similar values in their vector representations. In the field of NLP, a famous example of word embedding is King - Man + Woman = Queen, which can be interpreted as King is to Man as Queen is to Woman. In the example, if we compute the word embedding vector of $King - Man + Woman$, then the resulted vector should be at least approximately equal to the word vector of *Queen*. Such property suggests the word embedding can capture the meanings of and relationships between words (Mikolov et al. 2013b). This study used the word embedding from *bidirectional encoder representations from transformers* (BERT) to convert words into vectors. BERT was selected because its word embedding innovatively broke the limitations of previous work (Devlin et al. 2018). Traditional word embeddings such as *Word2Vec* (Mikolov et al. 2013a) cannot differentiate words with multiple meanings. For example, for the word *bank*, previous word embeddings cannot tell the difference of bank in “a bank to withdraw money” and “a river bank for a walk”. To ensure the evaluation of semantic coherence considers contextual meanings of words, we chose BERT’s word embedding in this study. Although BERT also uses the transformer architecture, the word vector extracted from it would not favor other transformer-based models. The first reason is that we are applying BERT’s word embedding to texts directly, the process of which are model agnostic. Second, previous studies have found that traditional machine learning models can achieve significant performance gains by using features extracted from BERT’s word embedding when dealing with text data for classification (Alimova and Tutubalina 2020; Roitero et al. 2020).

Cosine similarity is a measure of similarity between two vectors and is defined as:

$$similarity = \frac{A \cdot B}{\|A\|\|B\|}$$

, where the similarity is the value of *cosine* taking the angle between vector A and B. The value of cosine similarity ranges from −1 to 1. A pair of texts with a similarity of 1 suggests the pair is identical, and a similarity of −1 means the pair is completely opposite. In the study of Vakulenko et al. (2018), the researchers found a cosine similarity around 0.7 can best reflect texts’ strong coherence within a conversation. In our research, with the Python package *bert-as-service*, we computed the cosine similarity between original posts and generated texts from the models of RNN and GPT-2 to have an approximated coherence metric of generated texts.

Multidimensional Support from Texts by NLG Models

After evaluating models with quantitative measures, we further conducted a qualitative evaluation of the generated texts by the outperforming model. In the qualitative analysis, the authors manually examined the informational support, emotional support, and community support in original replies and generated replies with given posts. The two coders communicated actively and reached consensus on all the results. We rated the three types of support in original replies and generated replies with a scale of none (0), weak (1), moderate (2), and strong (3). A type of support coded with *none* meant the coders thought there was no such support found in texts. *Weak* support indicated that a type of support could be vaguely identified in texts, while such support was not the theme in texts. *Moderate* support suggested that a type of support could be explicitly identified in texts, while such support was not the theme in texts. *Strong* support meant that a type of support could be explicitly identified in texts, and the support served as the theme in texts. Examples can be found in Table 2. After the qualitative coding, we

Table 2 Examples of ratings on support with a mixture of original replies and generated replies

Support	Rating	Examples
Informational support	none	- “I think it is a great idea!”
	weak	- “I am still working on the details of the project. I am also interested in learning more about sustainable gardening. Any ideas on how to put this together? Thank you for your Interest.”
	moderate	- “Both are really good! The first one is cheerfull and the second is classy!:) I wouldn’t know which one to choose for now! Congrats!” - “Sounds great. I am also thinking of doing the same thing. Hope it helps. Here’s a link to my blog: [LINK]”
	strong	- “But the point is, these drugs have no quality or purity checks, and no guarantee that they are safe. They’re just there for people to be able to have a little fun, experiment, and create. And, as you said, they’re brewed up by inexperienced basement chemists…” - “Learning from mistakes and having the courage to make new mistakes is the most important skill to acquire in life. That’s a great question. I think we need to expand the definition of “mistake” to include failures in learning or personal development…”
Emotional support	none	- “:)”
	weak	- “Wow, that is really cool! Thanks! I’ll be sure to check it out. Thanks again!:)”
	moderate	- “Mistake also makes me grow like success does. Mistake was there whether you like it or not.” - “Sorry you have these roaming cats, they are such hunters.”
	strong	- “I also think that if you are passionate about something you can make it work. You just need to reform your strategy of working maybe and also it may take a long time until you find clientele to sell your creations for what they are worth. Don’t give up!” - “Yeah you don’t need to end it, there are many good options out there...Just stay focused on your goal, no matter how long you may have.”
Community support	none	- “The key thing we have to keep in mind is to provide value. A good way to do this is to make a directory of “good ideas” or “DIY””.
	weak	- “:)”
	moderate	- “Hi [NAME]! ‘Teaching how to fish’ is an expression that means that I will dedicate my time to teach a skill(s) to another person and I will take all what it needs to be done in order for the person to learn that skill...That is the idea!:)” - “Hi [NAME], I agree that Scrolling might make viewing easier. The problem is computer operation performance and the browser capability various…”
	strong	- “Hey [NAME], thanks for your comment. I totally agree with you…I’ll be sure to let you know how it goes.:)” - “Thanks for your feedback [NAME], I’ll try to come up with some ideas for a name soon.”

Note. Only essential content related to specific support has been presented. [NAME] refers to people’s names mentioned in the texts. [LINK] refers to web links shared in the texts conducted a MANOVA test to find if there existed a significant difference between the types of support offered by original replies (human learners) and generated replies. ANOVA tests were conducted to determine whether differences existed in each type of support between original and generated replies.

Before conducting the qualitative analysis, we first conducted a power analysis with *G*Power v3.1*. The power analysis for the MANOVA test suggested we needed a sample size of 120 in order to achieve a power of 0.95 at an alpha level of 0.05 and with a weak-to-medium effect size of 0.15. Therefore, 150 posts were randomly selected, with 300 replies in total (150 of original replies and 150 of generated replies).

Survey for NLG Model Validation

An experiment was then conducted to validate the findings from our previous quantitative and qualitative analysis. In the experiment, we recruited four upper-level graduate students with a background in educational technologies. There were two females and two males in the participants. Participants were asked to fill in a survey consisting of 10 post-reply pairs to answer (1) whether they think a reply was generated by a human or machine, (2) whether a reply provided informational, emotional, and community support, and (3) a 10-point Likert scale to suggest the quality of a reply in terms of grammars, readability, and coherence to the discussion context. Participants were allowed to select multiple support types for a reply if applicable or choose none if no evident support could be detected. Definition and examples of the three types of support were given to participants before the experiment.

Each survey had a bank of 15 post-reply pairs and each post had a human-generated and a machine-generated reply. Five human-generated and five machine-generated pairs were randomly selected for each participant. The participants were informed that the survey contained both human- and machine-generated replies, without knowing the exact number. To check the NLG model's generalizability in a similar course context, we randomly assigned the four participants to a group of 2 people to form two groups. One group used the bank of post-reply pairs from MOOC A, which was the course the NLG model was trained with. The other group used sources from another course of a similar topic on creative thinking (MOOC B), the data and context of which were new to the NLG model. Post-reply pairs from the two courses were extracted with a Python script, with only posts having at least one human-written reply being selected. Machine replies were then generated with the trained NLG model.

After participants submitted their responses, descriptive statistics were computed for further analysis. The analysis was conducted from two dimensions to compare the ratios of correctly labeled replies and socio-emotional support as well as the average ratings of content quality of replies. The two dimensions were (1) comparisons of performance between replies generated by the machine from the MOOC A group and the MOOC B group. This is to understand if the trained NLG model would yield a comparable performance in a similar while new context; and (2) comparisons of performance between replies generated by humans and the machine across all participants. This is to learn if the trained NLG model can generate replies similar to those generated by humans.

Results

Results for NLG Models with RNN and GPT2

The model training and evaluation were conducted with an NVIDIA Tesla P100 GPU of 16-gigabyte memory with Python 3. For the RNN model, we used the Python packages *Tensorflow* and *Keras*. In terms of GPT-2, we used the Python packages *PyTorch* and *Huggingface's Transformers*. Each model was stopped when training loss' decreasing tendency contradicted with that of validation loss, or when training loss stopped decreasing. The RNN model was stopped after fifteen epochs, and the GPT-2 model was stopped after eight epochs. An epoch of training means a full iteration of training data by the model since training with deep learning usually requires that data be chunked into mini-batches to avoid out-of-memory issues. Table 3 shows the evaluation results of RNN, GPT-2, and original replies from the perspectives of word perplexity, average F-K grade level, and average cosine similarity with posts.

The word perplexity of 37.94 of GPT-2 suggests the GPT-2 model is much more confident in predicting the next word given a context, compared to the word perplexity of 151.41 of the RNN model.

Compared with the RNN model's average F-K grade level of 22.68 ($SD = 22.01$), the GPT-2 model has an average F-K grade level of 10.10 ($SD = 13.40$). The average F-K grade level of GPT-2 is much closer to that of the original replies ($M = 8.73$, $SD = 9.20$). The GPT-2 model's distribution of the F-K grade level follows more closely with that of the original replies than the distribution of RNN (see Fig. 4). Through the inspection of randomly selected examples of different F-K grade levels (see Table 4), we find that texts' grade level scores falling in the range from 3 to 15 suggest better readability. In contrast, results with grade level higher than 20 or lower than 1 are usually repetitive or meaningless. From Fig. 4, F-K grade level scores of GPT-2 mainly lies in the range of 3 to 13, so are the scores of the original replies, while RNN has much more extreme scores. In general, the readability scores of GPT-2 better align with those of the original replies and GPT-2's readability scores are concentrated in the favorable range from observations. Therefore, we conclude that texts generated by the GPT-2 model have better readability than those of RNN.

A similar evaluation was conducted on cosine similarity. After evaluating texts on various cosine similarities, we find that generated replies with cosine similarity in the

Table 3 Evaluation results of RNN and GPT-2 compared with original replies

	Word perplexity	Avg F-K grade level	Avg cosine similarity
RNN	151.41	22.68	0.52
GPT-2	37.94	10.10	0.57
Original Replies	–	8.73	0.58

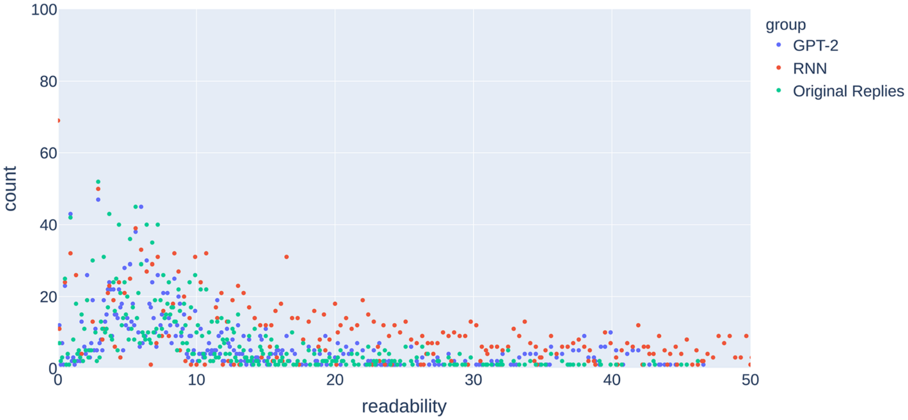

Fig. 4 F-K grade level of texts generated by GPT-2, RNN, and Original Replies

mid-upper range are best aligned with corresponding posts (see Table 5). In the results, the GPT-2 model has an average cosine similarity of 0.57 ($SD = 0.22$), with 90% of responses ($n = 2493$) falling in the range from 0.13 to 0.84, while the RNN model's mean cosine similarity is 0.52 ($SD = 0.17$), with 90% of responses ($n = 2493$) in the range from 0.17 to 0.75. The average cosine similarity of the original replies is 0.58 ($SD = 0.18$), with 90% of responses ($n = 2493$) falling in the range from 0.24 to 0.81. For responses with a cosine similarity greater than or equal to 0.95, there are 23 cases generated by GPT-2, none by the RNN model, and 3 in the original replies. Although GPT-2 tends to generate more responses with little variation from the original posts (cosine similarity >= 0.95), such cases only take less than 1% of the sample size ($N=$

Table 4 Examples of texts with different F-K grade levels

F-K grade level	Generated replies
0.1	"!!"
3.3	"We have a quad where we have to write a short poem. We write it and then we have to call it an essay. The students pick it up and have to fill in the blanks and sentences with the same words. It is a great exercise. I would love to see the result of your project."
7.7	"LOL. I wonder how many others have already posted their creative works here."
11.6	"Thank you! My project page has been working for me and has gotten tons of views so far, but I don't have enough people that I can invite to participate. I would love to find people willing to join and help me grow my project and make it successful! So, the question is, do you think it is unrealistic to have a goal of having kids stay home with the chores and do the housework? I know it sounds crazy but it actually may work out very well if you do!"
21.2	"I have a lot of ideas on how to get started. I am not sure how I would like to do this... but I've just started this, so I am going through the course to be able to add to the time a deadline. I found the that I had a thought that I was able to try it again."
31.7	"Silly thing for me to write. I am not sure if I am looking at the time to be able to do with my ". that's why I am doing. More than I like to do it. I have a week two things to get a handle on my own. I am a very open a person who is happier. Don' t worry. I' m from the northern california... I am a cousin, but I put in the middle of the past."

Table 5 Examples of texts with different cosine similarities

Cosine similarity	Post	Generated replies
0.1	"[NAME], [NAME], [NAME] and [NAME] and the rest - Beautiful work...I really can see the refinement in your logos with every feedback you received - I just stumbled upon this thread and I must say it is impressive."	:)
0.18	"I finished Dizzy and just have to clean up the punctuation, typos and add pictures. Plan to take some of my cat. He can be a model for me. The J.D. story is coming along, but the longer it gets, the harder it is for me to follow the choices. I have to find a way to end it. I'm getting lost in my own story."	"Thanks for a great thread."
0.35	"For anyone that's interested: I am working on a syllabus here: [LINK] I enabled commenting for everyone...so feel free to give me feedback as I go along. However, note that I just started so there is still a lot to do, and by no means am I close to the final project. Ultimately I have to run this by the professors tomorrow morning, and hopefully we can have this posted sometime tomorrow afternoon. Thanks everyone for your input."	"Thanks. I will keep you posted, it will be a fun process."
0.67	"I need a deeper explanation as well. I don't get what are projects actually and where are Project Reflection Surveys mentioned in 'About the course'"	"I'm really interested in find out more. Thanks a lot in advance. I'm sorry to hear that you were not able to get the "about the course" email. We are going to keep you posted on the progress of this project as well."
0.74	"Hi Everybody, I post what we have so far. In my opinion it can be now submitted by you… Title: Patria et Terra (Homeland and Earth) - intercultural film about our home Project Goals Mission Statement: Our goal is to create a video product in the theme of "Home" presented by different genres. The final product will consist of certain parts/scenes that are made by different people from different places in the world..."	"I would like to thank all contributors of ideas, knowledge and support. [NAME], I love this idea, it will be a great home movie to make in our native language. I will give you a call tomorrow to discuss the project."
0.85	"I am using the life ring exercise to get more focus in my life. Rather then determining my life ring in one go, I am developing it over time. So far I already eliminated some factors I thought absolutely important for me before. Let's where it leads me?"	"I'm not sure I'm going to follow the life ring exercise, but I'm going to try it."
1	"Hi [NAME], sorry for the late reply. Thank you very much for your mail and inspirational words! Yeah it does look like it has good song writing potential, thank you for the tip! Also thank you for the advice for my employment venture, I am going to use those	"Hi [NAME], sorry for the late reply. Thank you very much for your mail and inspirational words! Yeah it does look like it has good song writing potential, thank you for the tip! Also thank you for the advice for my employment venture, I am going to use those resources in my search as you have

Table 5 (continued)

Cosine similarity	Post	Generated replies
	resources in my search as you have suggested. Thanks again, [NAME]"	suggested. Thanks again, [NAME]"

2770). Therefore, the GPT-2 model is considered capable of generating more coherent replies than the RNN model, since the average cosine similarity of GPT-2 is closer to that of the original replies. Moreover, the majority of GPT-2's cosine similarities are in the favorable range found empirically. Figure 5 shows the distribution of the cosine similarity of GPT-2, RNN, and original replies.

Given the performance of GPT-2 excelled RNN from various perspectives, we chose GPT-2 over RNN for further analysis to compare the types of support from GPT-2 with those by human learners.

Results for Multidimensional Support from GPT-2

In addition to conducting MANOVA and ANOVA tests, we also calculated the percentage of supportive replies. A reply is deemed as supportive if the reply has a rating of moderate or strong support. Replies rated with none of or weak support are treated as non-supportive in that we think these replies only provide minimal support that might not be helpful for learners. Table 6 shows the descriptive statistics on the percentage of different support types, as well as the results of statistical tests. From the percentage of supportive replies, original and generated replies provide close numbers of emotional and community support. In contrast, the original replies have substantially more instances of informational support. In terms of the MANOVA test, a significant multivariate effect in supporting MOOC learners was found between original replies and generated replies, $\Lambda = 0.95$, $F(3, 296) = 4.85$, $p < 0.01$. However, the Wilk's Λ of 0.95 suggests only 5% of the variance can be explained by whether humans or machines generate a reply. The follow-up ANOVA tests do not find any significant

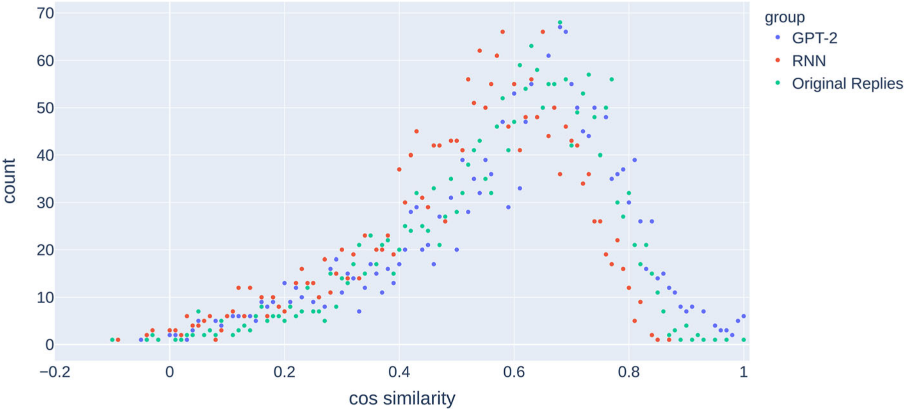

Fig. 5 Cosine similarity of texts generated by GPT-2, RNN, and Original Replies

Table 6 Statistical report on types of support between original and generated replies

	Percentage of supportive replies		MANOVA		ANOVA
	Original replies ($n = 150$)	Generated replies ($n = 150$)	Wilks' Λ	F	F
Informational support	42.85%	26.43%	–	–	11.73***
Emotional support	10.0%	8.57%	–	–	0.26
Community support	81.42%	79.28%	–	–	0.01
	–	–	.95	4.85**	–

Note. * means p value < .05, ** means p value < .01, and *** means p value < .001

differences in emotional and community support between original and generated replies. However, there was a significant difference in informational support between original replies and generated replies, $\eta^2 = 0.04$, $F(1, 298) = 11.725$, $p < 0.001$. Similarly, the small effect size (η^2) indicates that the effect of replies' generation sources on the provided informational support is trivial in practice.

Table 7 Descriptive statistics on the survey results to compare responses of GPT-2 model and Humans

		Group A (MOOC A)		Group B (MOOC B)	
Participants		P1	P2	P3	P4
# of replies labeled as machine		6	7	5	7
GPT-2-generated replies ($n = 5$)	# of correctly labeled	2	4	2	4
	% of correctly labeled	33.33	57.14	40.00	57.14
	# of informational support	4	3	3	3
	% of informational support	80.00	60.00	60.00	60.00
	# of emotional support	2	0	3	2
	% of emotional support	40.00	0.00	60.00	40.00
	# of community support	1	1	3	2
	% of community support	20.00	20.00	60.00	40.00
	Avg content quality	7	2.8	6.4	3.8
# of replies labeled as human		4	3	5	3
Human-generated replies (n = 5)	# of correctly labeled	1	2	2	2
	% of correctly labeled	25.00	66.67	40.00	66.67
	# of informational support	2	3	2	2
	% of informational support	40.00	60.00	40.00	40.00
	# of emotional support	2	0	1	2
	% of emotional support	40.00	0.00	20.00	40.00
	# of community support	2	2	2	3
	% of community support	40.00	40.00	40.00	60.00
	Avg content quality	6	6.2	4.6	6.4

Results for NLG Model Validation Surveys

Table 7 shows the descriptive statistics of each participant's result of human-machine labeling, social support detection, and average content quality. For example, participant 1 (P1) from Group A evaluated ten post-reply pairs from MOOC A. Among the ten pairs, six were labeled as "machine", while the rest four were thought to be created by humans. For the six replies marked as machine-generated, two of them were actually generated by the GPT-2 model, which had a 33.33% (2/6*100%) correction percentage. For the four replies marked as human-generated, only one was actually generated by humans, having a 25% (1/4*100%) correction percentage. In terms of social support, P1 thought four out of five (80%) machine-generated replies showed informational support, two out of five (40%) showed emotional support, and one out of five (20%) demonstrated community support. Regarding the content quality, the average value rated by P1 was 7 for the machine-generated replies ($n = 5$) and 6 for the human-generated replies ($n = 5$).

To understand the generalizability of the trained GPT-2 model, we descriptively compared the differences of label correction rate, social support detection rate, and ratings of content quality between machine-generated texts in Group A and B (see Fig. 6(1)). The results showed similar ratios of correct labels for machine-generated texts between Group A ($M = .4523$, $SD = .1684$) and Group B ($M = .4857$, $SD = .1212$. In the meantime, machine-generated texts received a higher ratio of informational support in Group A ($M = .7000$, $SD = .1414$) than Group B ($M = .6000$, $SD = 0$). On contrast, the ratios of emotional and community support of machine-generated texts from Group B ($M_{emotional} = .5000$, $SD_{emotional} = .1414$; $M_{community} = .5000$, $SD_{community} = .1414$) were higher than those from Group A ($M_{emotional} = .2000$, $SD_{emotional} = .2828$; $M_{community} = .2000$, $SD_{community} = 0$). A similar result was shown regarding the average ratings of content quality for Group A ($M = 4.90$, $SD = 2.97$) and Group B ($M = 5.10$, $SD = 1.84$). It was interesting to see that the numbers of emotional and community support from Group B were greater than those from Group A, even though the GPT-2 model was trained within the context of MOOC A. The findings suggest that the trained GPT-2 model's performance can be generalizable in a similar while new context.

To find out if the trained GPT-2 model can generate replies similar to those generated by humans, we descriptively examined the differences in terms of label correction rate, social support detection rate, and ratings of content quality between texts generated by the machine and humans (see Fig. 6(2)). Participants were almost

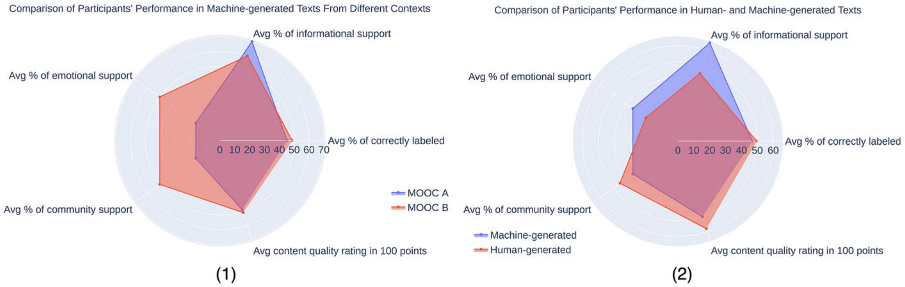

Fig. 6 Comparisons of participants' performance between human- and machine-generated texts and between MOOC A and MOOC B

equally likely to correctly label a human-generated reply ($M = .4958$, $SD = .2065$) and a machine-generated one ($M = .4691$, $SD = .1213$). Human-generated texts showed a higher ratio of community support ($M = .4500$, $SD = .1000$) than those generated by the machine ($M = .3500$, $SD = .1915$). However, higher ratios of informational ($M_{machine} = .6500$, $SD_{machine} = .1000$ vs. $M_{human} = .4500$, $SD_{human} = .1000$) and emotional support ($M_{machine} = .3500$, $SD_{machine} = .2517$ vs. $M_{human} = .2500$, $SD_{human} = .1915$) were found in machine generated texts. Lastly, the average content quality rated by participants was higher for human-generated texts ($M = 5.80$, $SD = .82$) than machine-generated texts ($M = 5.00$, $SD = 2.02$).

Discussion

A discussion forum is an essential place for MOOC learners where learning transactions happen when students provide support to each other (Almatrafi et al. 2018; Wang et al. 2015; Xing et al. 2019). However, students' communication in MOOC discussion forums is found to be scarce (Chiu and Hew 2018). The scarcity of interaction was also found in this study, with each post having only 0.6 replies on average. Furthermore, due to the size of MOOCs and the vast amount of posts generated every day, it is difficult for MOOC instructors to manually provide timely support for each learner (Almatrafi et al. 2018). This research aims to provide students with automatic support in MOOCs, more specifically, support students socio-emotionally by providing machine-generated replies for their posts in MOOC discussion forums in a more efficient way. The goal of this study was to examine the possibility of generating texts similar to those generated by humans with deep-learning-based NLG models. In this section, we discuss the results presented in the previous section.

Previous works on using NLG heavily rely on manual engineering to provide learning support (Demetriadis et al. 2018; Tegos et al. 2019), while such an approach requires advanced knowledge in algorithms and software engineering that are difficult for general researchers and educators. More importantly, NLG models with only manual engineering can limit the diversity of generated texts and topics. Therefore, NLG with deep learning shows potential in providing learners with automatic support in that deep learning does not require many formal specifications for computers from human operators to achieve desired results. This study has examined the use of state-of-the-art deep learning models GPT-2 and RNN to support students with text generation. Our results showed that NLG with deep learning could provide learners with genuine responses without digging deep into feature engineering. Specifically, the results showed that the GPT-2 model could generate more coherent and readable texts than the RNN model. In the observation, GPT-2 can generate texts naturally and contextually that are difficult to discern from human-written texts (see Table 2). The distributions of readability and similarity of GPT-2 are better aligned with those of the original replies than RNN (see Figs. 4 and 5). The main range of F-K grade level by GPT-2 (approximately 3 to 13) in this study resonates with Kincaid et al. (1975), where the researchers found the U.S. Navy's training manuals had an F-K grade level ranging from 5.5 to 16.3. While the F-K grade level has been widely used to evaluate content readability in different domains and is shown to be an effective metric (Feng et al. 2010; Lipari et al. 2019; Xu et al. 2016), there are concerns about it due to the weights of the word and sentence lengths in the calculation (McNamara et al. 2014). The advent of the internet has shaped language used online towards a shorter form with abbreviations and symbols (Tagliamonte and Denis 2008). The F-K grade level, in this case, might not always accurately show the content readability. For example, a sequence of repetitive symbols such as "!" will have a low F-K grade level (see Table 4). However, the meaning of these symbols is subject to diverse interpretations and is not highly readable. Future studies on automatic evaluation of readability might consider improved metrics such as Coh-Metric (McNamara et al. 2014). Coh-Metric is less sensitive to the lengths of words and sentences and considers the relationships between ideas in the text, which can be a more robust readability indicator than the F-K grade level. For the cosine similarity, 62% of responses generated by GPT-2 are in the range from 0.5 to 0.8, which is closer to the percentage of the original replies (66%) than that of RNN (58%). In a study by Vakulenko et al. (2018), the researchers found a cosine similarity around 0.7 can best reflect the texts' strong coherence within a conversation. Therefore, the main range of GPT-2's cosine similarity indicates a functional coherence and suggests the model can generate texts that will effectively address students' posts.

Social support theory has been studied extensively in online learning settings, and literature suggests informational, emotional, and community support are important to students' learning (Hew 2016; Tomkin and Charlevoix 2014; Xing et al. 2018). However, few works have examined the design and development of NLG models to support learning from the perspective of social support theory. This study aims to fill the gap. Thus, we have evaluated the content generated by human learners and the GPT-2 model in informational, emotional, and community support. A MANOVA and follow-up ANOVA tests were then run to find significant differences. Descriptive results show that human-generated texts tend to provide more informational, emotional, and community support. However, the lack of significance suggests the GPT-2 model can provide a similar extent of emotional and community support compared with human learners. While human learners' replies are significantly more informative, the small effect size suggests that such a difference between machine-generated replies is trivial in practice.

To confirm the findings from the qualitative analysis, we then applied the GPT-2 model with real users through surveys. From the descriptive statistics, participants were almost equally likely to correctly label a human-generated reply and a machine-generated one. In the meantime, participants tended to detect more diversified support from machine-generated texts than those generated by humans (see Fig. 6(2)). The results, to some extent, support our findings from the qualitative analysis that the GPT-2 model could generate texts similar to those generated by humans. Participants' similar label correction percentages on machine- (46.91%) and human-generated (49.58%) texts suggest that it was challenging for participants to tell who has generated the texts. The finding that the trained GPT-2 model could generate texts that were challenging to distinguish from human-written texts aligned with previous studies (Adelani et al. 2020; Zhang et al. 2020). In Adelani and his colleagues' study, GPT-2 was trained to generate product reviews for online commercial websites and participants were asked to differentiate between real and generated reviews. Their results showed that participants rated close fluency scores for real and generated reviews. Other than the ability to generate texts well imitating learners, we also found the GPT-2 model had the potential generalizability that could yield ideal performance in a similar but new context. The comparison between the two groups of participants showed close performances for machine-generated texts. In fact, participants found more social support from texts generated within the MOOC different from the training context (MOOC B) (see Fig. 6(1)).

The trained NLG model can be applied to MOOC contexts in various forms. For example, a web service can be programmed to provide application programming interfaces (APIs) for text generation in response to students' posts. A demo of such service can be viewed at https://youtu.be/CyKeyb7_XaA. In the demo, a Python script was written to scrape students' posts from MOOC A. Replies were generated with the trained GPT-2 model. The results in the demo suggested that the GPT-2 model can well respond to content not seen during the training. For example, for a post in 2019 that greeted everyone in the discussion forum, the GPT-2 model generated welcoming texts "sorry for the late reply. I am so excited to be in this course with you all!". The model continued as a teacher to give suggestions on how to connect with people in discussion forums. For another post in 2018, where the author complained about the assessment in the course, the GPT-2 model generated texts with empathy and encouraged the author to provide suggestions for improvements.

The demo was made simplistically to demonstrate the use of NLG models with deep learning in MOOC contexts. In a production environment, more complex systems can be built to streamline the support process of NLG models. For example, webhooks of MOOC platforms can be used to subscribe to new post creation, and replies can be generated for posts without replies within a given amount of time. A webhook is a system that can be configured to respond in conditioned circumstances with pre-programmed functions. Generated replies can also be automatically evaluated for quality control based on the F-K grade level and cosine similarity.

This study showed an exploratory work on using NLG with deep learning models to support MOOC learners. Future work may develop based on this study's methodology or to extend this study's research goal for improvements. For example, will the types of support offered by human replies and generated replies differ in other contexts such as discussion forums for online learning or MOOCs with technical topics? Or instead of providing support randomly or for posts without replies, will students benefit more in terms of informational, emotional, and community support if their posts are automatically classified as in urgent need? Future studies may also explore what forms of automatic textual support students would prefer to receive. For example, textual support can be generated directly on discussion forums, or the textual support can be given with a private web page.

Limitations

The study demonstrated the potential of using NLG with deep learning to provide students with automatic socio-emotional support. The capability to generate high-quality texts and the potential generalizability in new contexts of the deep learning model have demonstrated the possibility of achieving the target goal. However, there are a number of limitations, three of which are noted here.

The first limitation is that the current study has limited data from real-case applications. Although we have conducted a small-scale experiment to understand if the NLG model can generate texts similar to those generated by humans, more data is needed to offer sufficient statistical power for inference. Furthermore, collecting students' outcome data with the NLG model applied within a real MOOC environment might help us better understand the effects of machine-generated texts, which is not available in this study. Second, the data used in this study is from a cognitively less challenging environment than courses of topics such as engineering. Students mainly used the discussion forum to form groups and share ideas, and only a few posts were expressing negative emotions or seeking help. Therefore, the evaluation of informational and emotional support might not apply to more challenging MOOCs. Lastly, although it is helpful to evaluate the quality of generated texts with F-K grade level and cosine similarity, results cannot be guaranteed to be of high quality. Generated texts may have desired readability and coherence metrics while still being unreadable or out of context. The unpredictability of low-quality generated texts might yield negative experiences for MOOC learners.

Conclusion

By examining the texts generated by deep learning models RNN and GPT-2, this study conducted preliminary work to provide social support with automatically generated texts. The results suggested that the GPT-2 model greatly outperformed the RNN model in terms of word perplexity, readability, and coherence with contexts. Evaluation of the informational, emotional, and community support provided by human replies and GPT-2-generated replies indicated the generated texts by GPT-2 could provide a similar extent of emotional and community support as humans, while significantly less informational support was detected in machine-generated replies. Further survey analysis suggested that although differences exist between texts generated by the machine and humans in terms of easiness of differentiation, social support, and content quality, such differences were small and aligned with our prior results. The findings imply that it is possible to provide MOOC learners with automatic social support on a large scale. Ultimately, the goal is to support students by providing machine-generated replies for their posts in MOOC discussion forums so that there can be more interactions. As a result, it might lead to more engagement and lower dropout rates as students are socio-emotionally supported in a timely manner. The use of NLG with deep learning models also enriches the literature on providing learners with automatic support and extends the literature on supporting learning with NLG and deep learning.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

References

Adelani, D. I., Mai, H., Fang, F., Nguyen, H. H., Yamagishi, J., & Echizen, I. (2020). Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection. In *International Conference on Advanced Information Networking and Applications* (pp. 1341–1354). Cham: Springer.

Alimova, I., & Tutubalina, E. (2020). Multiple features for clinical relation extraction: A machine learning approach. *Journal of Biomedical Informatics, 103*, 103382.

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.

Babori, A., Zaid, A., & Fassi, H. F. (2019). Research on MOOCs in major referred journals. *The International Review of Research in Open and Distributed Learning, 20*, 3. https://doi.org/10.19173/irrodl.v20i4.4385.

Benamara, F., & Saint-Dizier, P. (2004). Advanced relaxation for cooperative question answering. In M. T. Maybury (Ed.), *New directions in question answering* (pp. 263–274). Menlo Park: AAAI Press.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research, 3*, 1137–1155 http://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf. Accessed 10 December 2019.

Budzianowski, P., & Vulić, I. (2019). Hello, It's GPT-2—How can I help you? Towards the use of pretrained language models for task-oriented dialogue systems. In A. Birch, A. Finch, H. Hayashi, I. Konstas, T. Luong, G. Neubig, Y. Oda, & K. Sudoh (Eds), *Proceedings of the 3rd Workshop on Neural Generation and Translation* (pp. 15–22). Association for Computational Linguistics. https://doi.org/10.18653/v1/d19-5602.

Caballé, S., & Conesa, J. (2019). Conversational Agents in Support for Collaborative Learning in MOOCs: An Analytical Review. In F. Xhafa, L. Barolli, & M. Greguš (Eds.), *Advances in Intelligent Networking and Collaborative Systems. INCoS 2018. Lecture notes on data engineering and communications technologies* (Vol. 23). Cham: Springer.

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.

Clark, E., Celikyilmaz, A., Smith, N. A. (2019). Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In A. Korhonen, D. Traum, & L. Màrquez (Eds), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2748–2760. Association for Computational Linguistics. doi:https://doi.org/10.18653/v1/P19-1264.

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.

Cutrona, C. E., & Russell, D. W. (1990). Type of social support and specific stress: Toward a theory of optimal matching. In B. R. Sarason, I. G. Sarason, & G. R. Pierce (Eds.), *Wiley series on personality processes. Social support: An interactional view* (pp. 319–366). New York: John Wiley & Sons. Retrieved 10 Jan 2020, from https://psycnet.apa.org/record/1990-97699-013.

Dale, R. (2016). The return of the chatbots. *Natural Language Engineering, 22*(5), 811–817. https://doi.org/10.1017/S1351324916000243.

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.

Demetriadis, S., Tegos, S., Psathas, G., Tsiatsos, T., Weinberger, A., Caballé, S., … Karakostas, A. (2018). Conversational Agents as Group-Teacher Interaction Mediators in MOOCs. *Proceedings of 2018 Learning With MOOCS (LWMOOCS)*, 43–46. https://doi.org/10.1109/lwmoocs.2018.8534686.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv, 1810*, *04805*.

Du, X., Shao, J., & Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. *ArXiv Preprint ArXiv, 1705*, 00106.

Dubosson, M., & Emad, S. (2015). The forum community, the Connectivist element of an xMOOC. *Universal Journal of Educational Research, 3*(10), 680–690. https://doi.org/10.13189/ujer.2015.031004.

Dufty, D. F., Graesser, A. C., Louwerse, M. M., McNamara, D. S. (2006). Assigning grade levels to textbooks: Is it just readability? *Proceedings of the Annual Meeting of the Cognitive Science Society, 28*, 1251–1256. Retrieved 10 Jan 2020, from https://escholarship.org/uc/item/44f184z9.

Evans, R., Piwek, P., Cahill, L. (2002). What is NLG? *Proceedings of the International Natural Language Generation Conference*, 144–151. Retrieved 21 Dec 2019, from https://www.aclweb.org/anthology/W02-2119.pdf.

Evermann, J., Rehse, J.-R., & Fettke, P. (2016). A deep learning approach for predicting process behaviour at runtime. In M. Dumas, & M. Fantinato (Eds), *Proceedings of International Conference on Business Process Management*, *Vol. 281* (pp. 327–338). Springer. https://doi.org/10.1007/978-3-319-58457-7_24.

Fang, A., Macdonald, C., Ounis, I., & Habel, P. (2016). Using word embedding to evaluate the coherence of topics from twitter data. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1057–1060. https://doi.org/10.1145/2911451.2914729.

Fedus, W., Goodfellow, I., & Dai, A. M. (2018). MaskGAN: Better text generation via filling in the_. *ArXiv Preprint ArXiv, 1801*, *07736*.

Feng, L., Jansche, M., Huenerfauth, M., Elhadad, N. (2010). A comparison of features for automatic readability assessment. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 276–284. Retrieved 24 Dec 2019, from https://dl.acm.org/doi/pdf/10.5555/1944566.1944598.

Ferschke, O., Yang, D., Tomar, G., Rosé, C. P. (2015). Positive impact of collaborative chat participation in an edX MOOC. In C. Conati, N. Heffernan, A. Mitrovic, & M. Verdejo (Eds), *Proceedings of International Conference on Artificial Intelligence in Education, Vol. 9112*, 115–124. Springer. https://doi.org/10.1007/978-3-319-19773-9_12.

Goldie, J. G. S. (2016). Connectivism: A knowledge learning theory for the digital age? *Medical Teacher, 38*(10), 1064–1069. https://doi.org/10.3109/0142159X.2016.1173661.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge: MIT press.

Graves, A. (2013). Generating sequences with recurrent neural networks. *ArXiv Preprint ArXiv, 1308*, 0850.

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.

House, J. S. (1981). *Work stress and social support*. Massachusetts: Addison-Wesley.

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). Generating natural language question-answer pairs from a knowledge graph using a RNN based question generation model. In M. Lapata, P. Blunsom, & A. Koller (Eds), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 376-385). Association for Computational Linguistics. https://doi.org/10.18653/v1/e17-1036.

Io, H. N., & Lee, C. B. (2017). Chatbots and conversational agents: *Proceedings of 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 215–219. https://doi.org/10.1109/ieem.2017.8289883.

Jung, Y., & Lee, J. (2018). Learning engagement and persistence in massive open online courses (MOOCS). *Computers & Education, 122*, 9–22. https://doi.org/10.1016/j.compedu.2018.02.013.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Institute for Simulation and Training, University of Central Florida. https://doi.org/10.21236/ada006655.

Kop, R., Fournier, H., & Mak, J. S. F. (2011). A pedagogy of abundance or a pedagogy to support human beings? Participant support on massive open online courses. *The International Review of Research in Open and Distributed Learning, 12*(7), 74–93. https://doi.org/10.19173/irrodl.v12i7.1041.

Kumar, R., & Rose, C. P. (2010). Architecture for building conversational agents that support collaborative learning. *IEEE Transactions on Learning Technologies, 4*(1), 21–34. https://doi.org/10.1109/TLT.2010.41.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *ArXiv Preprint ArXiv, 1909*, 11942.

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.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436–444. https://doi.org/10.1038/nature14539.

Lee, Y., & Choi, J. (2011). A review of online course dropout research: Implications for practice and future research. *Educational Technology Research and Development, 59*(5), 593–618. https://doi.org/10.1007/s11423-010-9177-y.

Lee, J. S., & Hsiang, J. (2019). Patent claim generation by fine-tuning openai GPT-2. *ArXiv preprint arXiv, 1907*, 02052.

Lin, C.-P., & Anol, B. (2008). Learning online social support: An investigation of network information technology based on UTAUT. *Cyberpsychology & Behavior, 11*(3), 268–272. https://doi.org/10.1089/cpb.2007.0057.

Lipari, M., Berlie, H., Saleh, Y., Hang, P., & Moser, L. (2019). Understandability, actionability, and readability of online patient education materials about diabetes mellitus. *American Journal of Health-System Pharmacy, 76*(3), 182–186.

Mackness, J., Waite, M., Roberts, G., & Lovegrove, E. (2013). Learning in a small, task–oriented, connectivist MOOC: Pedagogical issues and implications for higher education. *The International Review of Research in Open and Distributed Learning, 14*(4), 140–159.

Masters, K. (2011). A brief guide to understanding MOOCs. *The Internet Journal of Medical Education, 1*(2), 1–6. https://doi.org/10.5580/1f21.

McKeown, K. R. (1982). The TEXT system for natural language generation: An overview. *Proceedings of the 20th Annual Meeting on Association for Computational Linguistics*, 113–120. https://doi.org/10.3115/981251.981285.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., Cai, Z. (2014). Automated evaluation of text and discourse with Coh-Metrix. Cambridge University Press. https://doi.org/10.1017/cbo9780511894664.

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., Khudanpur, S. (2010). Recurrent neural network based language model. Paper presented at the *Eleventh Annual Conference of the International Speech Communication Association*. https://www.fit.vutbr.cz/research/groups/speech/publi/2010/mikolov_interspeech2010_IS100722.pdf. Accessed 10 December 2019.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J. (2013a). Distributed representations of words and phrases and their compositionality. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger (Eds), Advances in Neural Information Processing Systems: Vol. 26 (pp. 3111–3119). Retrieved 23 Dece 2019, from https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Mikolov, T., Yih, W. T., Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In R. Morante & S. Yih (Eds), Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies (pp. 746–751). Association for Computational Linguistics. https://doi.org/10.3115/v1/w14-1618.

Mittal, A., Vigentini, L., Djatmiko, M., Prusty, G., Sharma, Y., King, M. E. (2018). MOOC-O-Bot: Using Cognitive Technologies to Extend Knowledge Support in MOOCs. *Proceedings of the 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, 69–76. https://doi.org/10.1109/tale.2018.8615453.

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.

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.

Pereira, J. (2016). Leveraging chatbots to improve self-guided learning through conversational quizzes. In F. J. García-Peñalvo (Ed), *Proceedings of theFourth International Conference on Technological Ecosystems for Enhancing Multiculturality*, (911–918). Association for Computing Machinery. https://doi.org/10.1145/3012430.3012625.

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* (pp. 1–21). Knowledge and Learning: Technology. https://doi.org/10.1007/s10758-019-09414-9.

Pietquin, O., & Hastie, H. (2013). A survey on metrics for the evaluation of user simulations. *The Knowledge Engineering Review, 28*(1), 59–73. https://doi.org/10.1017/S0269888912000343.

Potash, P., Romanov, A., Rumshisky, A. (2015). Ghostwriter: Using an lstm for automatic rap lyric generation. In L. Màrquez, C. Callison-Burch, & J. Su (Eds), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1919–1924. Association for Computational Linguistics. https://doi.org/10.18653/v1/d15-1221.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9. Retrieved 28 Dec 2019, from http://www.persagen.com/files/misc/radford2019language.pdf.

Radziwill, N. M., & Benton, M. C. (2017). Evaluating quality of chatbots and intelligent conversational agents. *ArXiv Preprint ArXiv, 1704*, 04579.

Roitero, K., Bozzato, C., Mizzaro, V. D. M. S., Serra, G. (2020). Twitter goes to the Doctor: Detecting Medical Tweets using Machine Learning and BERT. In F. Couto & M. Krallinger (Eds), *Proceedings of the Workshop on Semantic Indexing and Information Retrieval for Health from heterogeneous content types and languages co-located with 42nd European Conference on Information Retrieval: Vol. 2619*. RWTH Aachen University. Retrieved 4 Dec 2019, from http://ceur-ws.org/Vol-2619/short1.pdf.

Ruey, S. (2010). A case study of constructivist instructional strategies for adult online learning. *British Journal of Educational Technology, 41*(5), 706–720. https://doi.org/10.1111/j.1467-8535.2009.00965.x.

Serban, I., Sordoni, A., Bengio, Y., Courville, A., & Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. *Proceedings of the AAAI Conference on Artificial*

Intelligence, 30(1), 3776–3783. Retrieved 2 Jan 2020, from https://dl.acm.org/doi/10.5555/3016387.3016435.

Shah, D. (2019). Online degrees slowdown: A Review of MOOC Stats and Trends in 2019 — Class central. Class central Retrieved 4 Jan 2020, from https://www.classcentral.com/report/moocs-stats-and-trends-2019.

Shawar, B. A., & Atwell, E. (2007a). Different measurements metrics to evaluate a chatbot system. *Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*, 89–96. https://doi.org/10.3115/1556328.1556341.

Shawar, B. A., & Atwell, E. (2007b). Chatbots: Are they really useful? *Journal for Language Technology and Computational Linguistics, 22*(1), 29–49.

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.

Silverman, T. (1999). The internet and relational theory. *American Psychologist, 54*, 780–781. https://doi.org/10.1037/0003-066X.54.9.780.

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., et al. (2015). A neural network approach to context-sensitive generation of conversational responses. *ArXiv Preprint ArXiv, 1506*, 06714.

Sunar, A. S., White, S., Abdullah, N. A., & Davis, H. C. (2016). How learners' interactions sustain engagement: A MOOC case study. *IEEE Transactions on Learning Technologies, 10*(4), 475–487. https://doi.org/10.1109/TLT.2016.2633268.

Tagliamonte, S. A., & Denis, D. (2008). Linguistic ruin? LOL! Instant messaging and teen language. *American speech, 83*(1), 3–34.

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

Tegos, S., Psathas, G., Tsiatsos, T., Demetriadis, S. N. (2019). Designing Conversational Agent Interventions that Support Collaborative Chat Activities in MOOCs. In M. Calise, C. Kloos, C. Mongenet, J. Reich, J. Ruipérez-Valiente, G. Shimshon, T. Staubitz, & M. Wirsing (Eds), *Proceedings of Proceedings of Work in Progress Papers of the Research, Experience and Business Tracks at EMOOCs 2019: Vol. 2356* (pp. 66-71). Retrieved 20 Dec 2019, from http://ceur-ws.org/Vol-2356/research_short11.pdf.

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. *Proceedings of the First ACM Conference on Learning@ Scale Conference*, 71–78. https://doi.org/10.1145/2556325.2566245.

Tsai, H. S., Shillair, R., & Cotten, S. R. (2017). Social support and "playing around" an examination of how older adults acquire digital literacy with tablet computers. *Journal of Applied Gerontology, 36*(1), 29–55. https://doi.org/10.1177/0733464815609440.

Vakulenko, S., de Rijke, M., Cochez, M., Savenkov, V., & Polleres, A. (2018). Measuring semantic coherence of a conversation. In *International semantic web conference* (pp. *634–651*). Cham: Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds), *Advances in Neural Information Processing Systems: Vol. 30* (pp. 5998–6008). Retrieved 27 Dec 2019, from https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wang, X., Yang, D., Wen, M., Koedinger, K., Rosé, C. P. (2015). Investigating how Student's cognitive behavior in MOOC discussion forums affect learning gains. In O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds), *Proceedings of the 8th International Conference on Educational Data Mining* (pp. 226–233). International Educational Data Mining Society. Retrieved 2 Jan 2020, from https://files.eric.ed.gov/fulltext/ED560568.pdf.

Wang, W., Gan, Z., Wang, W., Shen, D., Huang, J., Ping, W., ... & Carin, L. (2018). Topic compositional neural language model. In A. Storkey & F. Perez-Cruz (Eds), *Proceedings of Machine Learning Research: Vol. 84. International Conference on Artificial Intelligence and Statistics* (pp. 356-365). Retrieved 23 Dec 2019, from http://proceedings.mlr.press/v84/wang18a/wang18a.pdf.

Wen, M., Yang, D., Rose, C. (2014). Sentiment analysis in MOOC discussion forums: What does it tell us? In J. Stamper, Z. Pardos, M. Mavrikis, & B.M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 130-137). Retrieved 2 Jan 2020, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.660.5804&rep=rep1&type=pdf.

Wen, T. H., Gašic, M., Mrkšic, N., Rojas-Barahona, L. M., Su, P. H., Vandyke, D., Young, S. (2015). Toward multi-domain language generation using recurrent neural networks. In K. Knight, A. Nenkova, & O. Rambow (Eds), *Proceedings of the 2016 Conference of the North American Chapter of the Association* *for Computational Linguistics: Human Language Technologies* (pp. 120-129). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1015.

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). Thousand Oaks: Sage Publications, Inc..

Wise, A. F., Cui, Y., Jin, W., & Vytasek, J. (2017). Mining for gold: Identifying content-related MOOC discussion threads across domains through linguistic modeling. *The Internet and Higher Education, 32*, 11–28. https://doi.org/10.1016/j.iheduc.2016.08.001.

Wu, M., Xu, X., Kang, L., Zhao, J. L., & Liang, L. (2019). Encouraging people to embrace feedback-seeking in online learning: An investigation of informational and relational drivers. *Internet Research, 29*(4), 749–771. https://doi.org/10.1108/IntR-04-2017-0162.

Xing, W. (2019). Exploring the influences of MOOC design features on student performance and persistence. *Distance Education, 40*(1), 98–113. https://doi.org/10.1080/01587919.2018.1553560.

Xing, W., & Du, D. (2019). Dropout prediction in MOOCs: Using deep learning for personalized intervention. *Journal of Educational Computing Research, 57*(3), 547–570. https://doi.org/10.1177/0735633118757015.

Xing, W., Chen, X., Stein, J., & Marcinkowski, M. (2016). Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Computers inhuman behavior, 58,* 119–129. https://doi.org/10.1016/j.chb.2015.12.007.

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, W., Napoles, C., Pavlick, E., Chen, Q., & Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics, 4*, 401–415. https://doi.org/10.1162/tacl_a_00107.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized autoregressive Pretraining for language understanding. *ArXiv Preprint ArXiv, 1906*, 08237.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., … Dolan, B. (2020). DialoGPT: Large-scale generative pre-training for conversational response generation. In A. Celikyilmaz, T. Wen (Eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 270–278). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.30.

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.