Detecting AI-generated essays: the ChatGPT challenge

Ilker Cingillioglu
Department of Business Information Systems, The University of Sydney, Sydney, Australia

Received 25 March 2023
Revised 9 April 2023
Accepted 9 April 2023

Abstract

Purpose – With the advent of ChatGPT, a sophisticated generative artificial intelligence (AI) tool, maintaining academic integrity in all educational settings has recently become a challenge for educators. This paper discusses a method and necessary strategies to confront this challenge.

Design/methodology/approach – In this study, a language model was defined to achieve high accuracy in distinguishing ChatGPT-generated essays from human written essays with a particular focus on "not falsely" classifying genuinely human-written essays as AI-generated (Negative).

Findings – Via support vector machine (SVM) algorithm 100% accuracy was recorded for identifying human generated essays. The author discussed the key use of Recall and F2 score for measuring classification performance and the importance of eliminating False Negatives and making sure that no actual human generated essays are incorrectly classified as AI generated. The results of the proposed model's classification algorithms were compared to those of AI-generated text detection software developed by OpenAI, GPTZero and Copyleaks.

Practical implications – AI-generated essays submitted by students can be detected by teachers and educational designers using the proposed language model and machine learning (ML) classifier at a high accuracy. Human (student)-generated essays can and must be correctly identified with 100% accuracy even if the overall classification accuracy performance is slightly reduced.

Originality/value – This is the first and only study that used an n-gram bag-of-words (BOWs) discrepancy language model as input for a classifier to make such prediction and compared the classification results of other AI-generated text detection software in an empirical way.

Keywords AI generated Text detection, ChatGPT, OpenAI, Machine learning, Recall, F2 score

Paper type Research paper

Introduction

Artificial intelligence (AI)-generated text, also known as computer- or machine-generated text, refers to the output of text produced by a computer or machine using natural language processing (NLP) and machine learning (ML) techniques. AI has become sophisticatedly "intelligent" and prolific enough to produce highly original and quality essays posing an unprecedented threat to the integrity of studentship and academia if left unchecked and undetected. Developed and released by OpenAI, ChatGPT has gone viral and become the latest craze in the world of generative AI since the end of 2022. Having attracted billions of dollars from tech investors such as Microsoft, OpenAI was valued at $29 billion by January 2023 (Saul, 2023).

It is no secret that students can easily push a button and the AI writes a paper for them (Turnitin, 2020). There is also no doubt that writing essays with AI is equivalent to plagiarism and considered academic cheating (Fyfe, 2022). Yeadon *et al.* (2023) found that plagiarism detection software such as Turnitin and Grammarly returned low plagiarism scores (1–2%) on ChatGPT (AI) generated short essays. In addition, some believe that deploying software that differentiates AI-generated essays from student-written essays is a futile exercise (Sharples, 2022). Although universities merely look for plagiarism currently rather than whether essays were AI-generated (Yeadon *et al.*, 2023), soon they will realize the serious threat such practise pose to academic integrity hence their reputation.

The International Journal of Information and Learning Technology
Vol. 40 No. 3, 2023
pp. 259-268

2056-4880
DOI 10.1108/IJILT-03-2023-0043

The next section provides an overview of the current literature on the applications and methods of AI-generated text, focusing on detecting AI-generated essays and challenges of this technology. Then, we present two excerpts as a sample case to showcase the competency of ChatGPT as we demonstrate the severity of the major challenge at hand. Afterward, we define a language model to meet the challenge of achieving high accuracy in distinguishing ChatGPT-generated essays from human written essays, report the results of the model, discuss the importance of Recall and F2 scores in measuring the proposed ML classifier's performance and finally compare the results to those of AI-generated text detection software developed by OpenAI, GPTZero and Copyleaks.

Background

In 1966, ELIZA, the first known chatbot which used pattern matching to simulate a conversation with a therapist, made natural language communication between human and machine possible (Weizenbaum, 1966). A few years later in 1970 came Terry Winograd's SHRDLU which was able to understand natural language input and generate text-based responses (Winograd, 1970). However, both ELIZA and SHRDLU were inherently limited to a specific domain as they did not have the ability to generate free-form text.

In the 1980s, researchers developed statistical language models that could predict the likelihood of a given word appearing in a sequence of text and developed models of discourse strategies to improve natural language generation (NLG) process (McKeown, 1985). These models were able to generate text by selecting the most likely word at each step in the generation process. One notable development in the 1990s was the introduction of Hidden Markov Models (HMMs) which were statistical models that capture the probability distribution over sequences of words (Eddy, 1998), making them useful for tasks such as part-of-speech tagging (Kupiec, 1992) and language modeling in conjunction with rule-based systems for speech and language processing (Knill and Young, 1997). Another notable breakthrough in the 1990s was the development of knowledge-based systems for text generation. These systems relied on ontologies and expert systems to generate textual content based on pre-defined rules and domain-specific knowledge. One example was a chatbot, released by Rollo Carpenter in 1997, called Cleverbot which would learn from the inputs provided by users through its own feedback loop (Love, 2014). These systems predominantly used hand-coded grammars and semantic rules to generate text that was to some extent syntactically and semantically accurate. However, these systems in 1990s were often inflexible, required substantial effort to develop and limited by inadequate computing power and scarcity of large-scale datasets.

In the 2000s, there was a resurgence of interest in statistical language models, particularly in the area of neural language models (Kumar and Thakur, 2012). These models use deep neural networks (NN) to learn the statistical structure of language and generate text based on this learned structure (Hermans and Schrauwen, 2013). This approach has led to significant advances in AI text generation, including the development and evolution of OpenAI's generative language models such as generative pre-trained transformer (GPT)-1 (trained on a dataset of nearly 8 million web pages and had 117 million parameters), GPT-2 (trained on a larger dataset of around 40 GB of text and had 1.5 billion parameters) and GPT-3 (trained on a much larger dataset of around 570 GB of text and had 175 billion parameters) which can generate highly relevant and coherent text that is hardly distinguishable from human generated text (Brown, 2023).

AI-based text generation has come a long way in recent years, from simple rule-based systems to advanced NN-based models capable of producing human-like text. The utility of deep learning models and NLP techniques has revolutionized text generation and allowed for greater control and accuracy over the output (Dwivedi *et al.*, 2023). With further advancements in AI technology, text generation is set to become even more sophisticated and useful in a myriad of applications.

ChatGPT-4

One of the fastest growing applications of AI-created text is in NLG, which is a software process of generating text or speech from non-linguistic input (Gatt and Krahmer, 2018). NLG is used in a variety of domains, including but not limited to content creation for websites and social media, news generation and development of chatbots to name a few.

The chatbot of OpenAI, ChatGPT-4 (or ChatGPT: Chat Generative Pre-trained Transformer) uses a technique called "autoregression" along a more advanced NLG method called neural text generation through a model called transformer-based neural network (T-NN or T-Net) to generate human-like text (OpenAI, 2023b). Autoregression is a form of self-supervised ML where the T-NN model is trained to predict the next word in a sentence, given the previous words. It is a type of language model that is trained on a large dataset of text, mostly acquired from books, articles and websites. The model learns patterns and structure in the data, such as grammar, vocabulary and common phrase structures. To generate text, the model starts with an initial prompt, such as a few words or a sentence and then the next word in the sequence is generated. Until the last word or phrase in the sequence is generated, this process is continued one word at a time. The deep NN with 175 billion parameters allows GPT3 to generate text that is logical, relevant, coherent and grammatically accurate (OpenAI, 2023b).

One of the main approaches for detecting AI-generated text is ML-based, where a classifier is trained to recognize patterns in the text that feature computer-generated text. This can be achieved using a variety of algorithms, such as support vector machine (SVM), decision trees, random forest (RF) and deep learning methods including recurrent neural networks (RNNs) and convolutional neural networks (CNNs). Another method to detect AI-generated text is based on linguistic characteristics of text indicating cues of machine generation. Examples of such characteristics include patterns in punctuation and sentence length, n-gram frequency and the use of particular words or phrases (Heikkilä, 2022).

Since ChatGPT keeps evolving and already produces high-quality human-like output, existing detection models hardly keep up [8]. Therefore, although many methods have been proposed for detecting AI-generated text, their results do not apply to what the latest version of ChatGPT can produce. It's not a matter of if but when a student submits an essay generated by ChatGPT, academics need an AI-led detection tool which also keeps evolving to keep up with the newly acquired capabilities of advanced AI-driven NLG tools like ChatGPT. The greatest challenge, however, is the accuracy of detection. More importantly, genuinely human-generated essays should not be falsely classified as AI-generated for ethical, legal and professional reasons. This paper aims to address these challenges by proposing an AI-led detection method and a prediction (i.e. made by an ML classifier) performance measure.

Sample case

In an experiment, teachers could not differentiate AI-generated essays from human-written essays (Olsson and Engelbrektsson, 2022). Here we present two excerpts (Table 1) to demonstrate how competently and indistinguishably ChatGPT can generate an essay like a human (i.e. student). One of them was written by the author of this study when he was a grad student at the University of Melbourne, while the other excerpt by ChatGPT. Both excerpts address the same essay question about data-driven predictive policing. Can you tell which one was written by the author (i.e. human) ? More importantly, if a student submits the essay that you suspect was generated by AI, what can you do to back up your suspicion?

The language model

Bag-of-word (BOW) language models are commonly used in NLP tasks such as language translation, sentiment analysis and text classification. The basic idea behind BOW models is

Excerpt 1	Excerpt 2
Data-driven predictive policing uses algorithms and data analysis to identify crime patterns and predict where crimes are likely to occur in the future. This information is then used to target policing efforts in those areas, with the goal of preventing crime before it happens. One of the benefits of predictive policing is that it can help law enforcement agencies make more informed decisions about how to allocate resources. By identifying crime hotspots and patterns, agencies can target their efforts in areas where they are most needed, potentially leading to a reduction in crime. Additionally, predictive policing can help law enforcement agencies identify patterns and trends that may not be immediately apparent, such as the presence of a specific criminal organization or a rise in a specific type of crime. However, there are also significant drawbacks to this approach. One concern is that predictive policing can perpetuate existing biases in the criminal justice system. If the data used to make predictions is not representative of the entire population, it can lead to over-policing in certain neighborhoods, disproportionately impacting communities of color, low-income neighborhoods and other marginalized groups. Additionally, if the algorithm is trained on data that reflects historical biases, it can perpetuate these biases and lead to discrimination in policing practices. Another drawback is that it can lead to over-policing in certain neighborhoods, which can result in a negative impact on community relations and lead to mistrust of law enforcement. This can lead to increased tensions between the community and the police and make it more difficult for the police to effectively do their job. In order to address these concerns, it's important that predictive policing is used in conjunction with community engagement and oversight. This can include involving community members in the development and implementation of predictive policing programs, as well as regularly evaluating and adjusting the algorithms to ensure they do not perpetuate bias. Additionally, transparency and explainability of the algorithms used in predictive policing can help to increase public trust and understanding of the technology. In conclusion, predictive policing has the potential to be a valuable tool for reducing crime and making policing more efficient, but it must be used with caution and be paired with community engagement and oversight to ensure that it does not perpetuate bias or discrimination. Additionally, regular evaluations and adjustments of the algorithms used in predictive policing are crucial to ensure that they are fair, accurate and unbiased	Data-driven predictive policing incorporates analyzing past crime data and then allocating resources in response to that crime data and making sure that police officers are in the right places (hotspots), at the right times. The effectiveness of the DDPP applications hinges predominantly on the quality of data that feeds into it. For now, its algorithms locate hotspots based on previously reported crimes; however, reported crime data may not always be true. Moreover, predictive policing can drive blind justice which certainly may not be the best kind of justice. For example, when police officers are sent to the hotspots, they may begin to feel like they are more entitled to do searches and stops in those hotspots than they otherwise would be and that judges and juries will be more accepting of the justifications for over-policing. Researchers and software developers have been working on projects to take predictive policing to the next level by creating algorithms that profile individuals based on their prior conducts and social media activities so that enforcement agencies can take preventative measures before some potential individuals commit crimes. However, these projects raise ethical concerns about predictive policing entrenching retrospective discrimination and bias into future policing decision-making and strategies. Understanding the limits and implications of each step of collecting, analysing and utilizing crime data is paramount for acknowledging how predictive policing technologies – regardless of being considered productive, unbiased and non-discriminatory – may inherit the prejudice and bias of its collectors, analysts and end-users. Although there is some evidence of improved productivity in crime reduction through predictive policing, individuals involved in the justice and enforcement system need to be educated in regard to recognizing its limits and offsetting its drawbacks. To conclude, although predictive policing technologies can offer productive ways to reduce crime, innocent people's past or even social media activity can make them a target for the police. Since the ends should not justify the means, making people a target due to their past or online behavior for the sake of allocating police resources more efficiently to fight crime is morally wrong because it not only violates the privacy, civil and human rights of the targeted people but also raises unfair, discriminatory and biased over-policing concerns towards mostly underprivileged communities. After all, all people should be subject to equal justice and be treated in a fair and equitable manner no matter where they live, what they have done in the past and what they read or share online

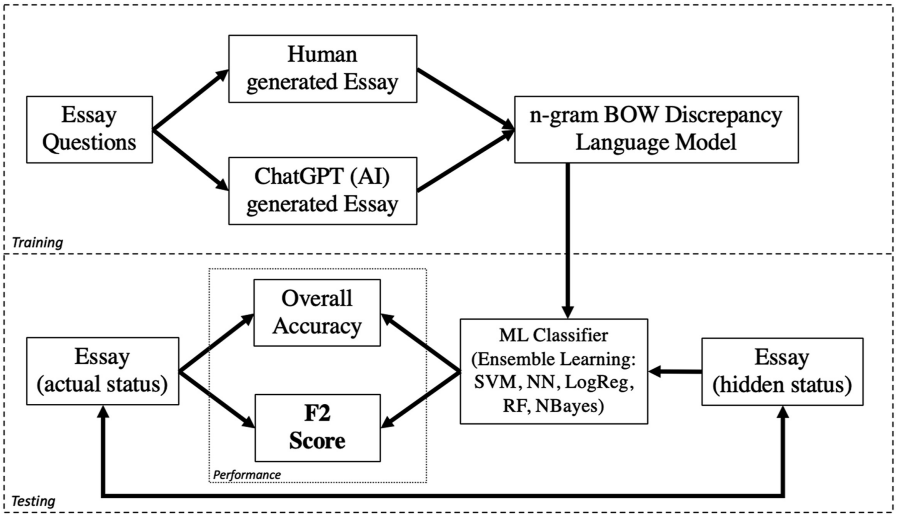
Source(s): Authors' own creation

Table 1.
Two sample excerpts. One written by a former student and the other by ChatGPT on data driven predictive policing

to represent a piece of text as a "bag" (or set) of its words, disregarding grammar and word order but keeping track of the frequency of each word. In other words, BOW models are a type of feature representation for text data. They are simple to implement as they represent a piece of text as a fixed-length vector, where each element of the vector corresponds to a word in a predefined vocabulary and the value of the element is the "frequency" (or count) of that word in the text.

We represented each text as BOW vector in the essays generated by humans (i.e. students) and by ChatGPT separately. Standard BOW models don't consider the order of words or grammatical structure of sentences. So, we developed an n-gram version which considers up to 5 consecutive words (n = 5). Since the main purpose of the ML classifier to be built upon the model is to differentiate human generated essays from AI-generated ones, we designed the language model in a way that the vectoral differences between human and ChatGPT generated essays are calculated. We called it "n-gram Discrepancy BOW language model" (Figure 1).

Where n is a positive integer, an n-gram is a contiguous sequence of n items from a given sample of text. In the context of BOW models, n-grams can be used as a way to integrate some information about word order and context into the feature representation. Instead of considering each word individually, n-gram models consider groups of words together, where each group is called an n-gram. For instance, in a bigram model (where n = 2), the text " all people should be subject to equal justice" would be represented as the set of bigrams {"all people", "people should", "should be", "be subject", "subject to", "to equal", "equal justice"}. Likewise, in a trigram model (where n = 3), the text would be represented as the set of trigrams {" all people should", " people should be", " should be subject", "be subject to", "subject to equal", "to equal justice"}. When creating the BOW representation, the frequency of each n-gram is counted and used as the value of the corresponding element in the vector. N-gram models can be useful for capturing more contexts about the words and the order in which they appear in the text, which can be helpful for an NLP task such as classifying human and AI-generated text. However, it should be noted that the dimensionality of the feature space increases exponentially with the value of n, which can make the model



Source(s): Authors own creation

Figure 1. Training the n-gram Discrepancy Bag-of-Words (BOWs) language model with Human (student) and ChatGPT (AI) generated essays and Testing the Machine Learning (ML) Classifier yielding Overall Accuracy and F2 Score

unreasonably complex and computationally costly. Therefore, when using n-grams in BOW models, it's important to strike a balance between the computational complexity of the model and the amount of context captured.

We procured 115 student essays and asked the same questions to Copyleaks (2023, March 14 Version) until we received 115 AI-generated responses (recorded as AI-generated essays). We ensured that every essay (AI or human generated) we included in the study was made of minimum 400 words. Upon a total number of 230 essays ($N = 230$), the training of the BOW model predominantly incorporated identifying the n-gram differences (i.e. discrepancies) between human written and ChatGPT generated essays. Accordingly, the ML classifier was built based on the spread and diversity of the dynamic n-gram discrepancies in the BOW model. Since ChatGPT is a sophisticated generative pre-trained transformer with an AI evolving and becoming more and more competent at generating natural language (human-like) text, the detection system must constantly keep evaluating the discrepancies between human generated and AI-generated essays. To do so, the n-gram discrepancy BOW model was continually trained with new essays written by humans and ChatGPT and the classification algorithms were updated accordingly.

In the testing phase, the data collected and vectorized via the BOW model are transferred to an ML classifier which in addition to the word count of each essay takes into account the spread and diversity of the n-gram differences between human and AI-generated essays. Upon a 70/30 data partition and 10-fold Monte Carlo cross validation, we tested the performance of several classification algorithms such as Support Vector Machine (SVM), Naïve Bayes (NBayes), Logistic Regression (LogReg), Random Forest (RF) and Neural Network (NN). We also tested the performance of an Ensemble Learning (EL) classifier comprising all these 5 algorithms via majority voting. We modified the cut-off point of SVM (since it is the most discriminating algorithm) to eliminate False negatives entirely, hence we could achieve a perfect Recall score (1: 100%) and boost F2 score which in our case (i.e. differentiating student essays from AI-generated essays) should be considered more important than the overall accuracy.

F2 score

The F2 score is one of the measures of an ML model's accuracy. It puts a heavier emphasis on recall, as a greater weight is assigned to recall than precision. It ranges from 0 to 1, with a higher score indicating better accuracy. The F2 score is usually used when the data is imbalanced and/or when the cost of False Negatives is high.

False Negatives (FNs) in our confusion matrix indicate the number of actual human generated essays incorrectly identified as AI-generated and False Positives (FPs) indicate the number of actual AI-generated essays incorrectly identified as human generated (Table 2 and Table 3). Since the essays that are falsely classified but in fact genuinely written by students (i.e. FNs) will be subject to some sort of punishment, eliminating FNs should be a top priority in AI-generated essay detection.

Estimating merely the Accuracy (TP + TN/All values) of the ML classifier will not provide sufficient information about FPs and more importantly FNs. Therefore, we need to use

Table 2. Confusion matrix descriptions used for ML model performance evaluation

True positive (TP)	Correctly identifying human generated essays as human generated
True Negative (TN)	Correctly identifying AI-generated essays as AI-generated
False Positive (FP)	Incorrectly identifying AI-generated essays as human generated
False Negative (FN)	Incorrectly identifying human generated essays as AI-generated

Source(s): Authors' own creation

relevant metrics that consider the effect of FPs and FNs. One of these metrics is Precision (TP/(TP + FP)) and the other Recall (TP/(TP + FN)). Striking a balance, however, between Recall and Precision through F1 scoring is not a sensible strategy because eliminating the predictions of actual human generated essays falsely identified as AI-generated (FNs) should be considered more important than eliminating the predictions of actual AI-generated essays falsely identified as human generated (FPs). The discrepancy in the degree of importance can be met by assigning distinct weights to recall and precision. Specifically, when estimating the performance of our ML model, to do so, we must prioritize minimizing FNs over FPs by assigning more weight to FNs than FPs. It's worth noting that in such cases, the F-beta score is used, where beta (β) is a parameter that can be set to control the balance between precision and recall. The reasons for deciding exactly how much more weight (β = 1.5, or 2, 3, . . .) to allocate is out of the scope of this study. In the end, we gave twice the weight (β = 2) to Recall versus Precision. The formulas for the Fβ and hence F2-score are:

$$F_\beta\, score = \left(1+\beta^2\right) * \frac{Precision * Recall}{\left(\beta^2 * Precision\right) + Recall} \Rightarrow$$

$$F_2\, score = 5 * \frac{Precision * Recall}{(4 * Precision) + Recall}$$

Where Precision is the number of True Positive predictions divided by the number of True Positive predictions plus False Positive predictions and Recall is the number of True Positive predictions divided by the number of True Positive predictions plus False Negative predictions. In the case of F2, beta is set to 2.

Although overall accuracy in differentiating human written essays from AI-generated ones is crucial, the SVM estimation algorithm of our proposed model puts more emphasis on eliminating False Negatives (FNs) than it does for eliminating False Positives (FPs). After all, we don't want any genuinely written essay to be falsely classified as AI-generated (the other way around is to some extent acceptable). So, keeping FNs at 0 and attaining a perfect Recall score (1: 100%) should be a priority due to our ethical and professional code of conduct to uphold the presumption of innocence. This also translates to if an essay is classified as AI-generated (Negative), the verdict should desirably be 100% true. However, due to trade-off, we must accept that some AI-generated essays might be falsely classified as human generated (FPs).

Other detection software

Using a new sample (collected the same way as the training sample) of 150 essays (75: Human-generated; 75: AI-generated) we compared the output of our proposed model's classification algorithms to those of AI-generated text detection software developed by OpenAI (2023b), GPTZero 92023) and Copyleaks (2023).

Predicted \ actual	AI-generated	Human-generated
AI-generated	TN	FN
Human-generated	FP	TP

Note(s): AI-generated essays are classified as Negative, while Human-generated essays as Positive. TN: True Negative, FN: False Negative, FP: False Positive and TP: True Positive
Source(s): Authors' own creation

Table 3. Structure of the 2 × 2 confusion matrix used for ML model performance evaluation

Although, in terms of overall accuracy OpenAI (96.7%), GPTZero (96%) and Copyleaks (95.3%) were superior to our best performing models, none of these three detectors were flawless in terms of correctly classifying all human-written essays as OpenAI text detector yielded 2, GPTZero 3 and Copyleaks 5 FNs with 97.3%, 96 and 93.3% Recall scores respectively (Table 4).

Similar to our ML testing, in this new sample we achieved the perfect Recall score (1: 100%) and 0 FNs via SVM algorithm. The overall accuracy was 92.7%. So, in this case, while there is a 7.3% chance that an AI-generated essay be falsely classified as human-generated, none of the human-generated essays were falsely classified as AI-generated. Although the exact Blackstone ratio (10:1) is not pertinent, we believe that it is better to let multiple culpable persons escape than to make one innocent suffer. We also note that the EL yielded a better overall accuracy (94%) than the SVM (92.7%). However, in terms of recall and F2 score it was inferior to SVM. More importantly, SVM produced the best F2 score (97.2%) among all algorithms including the ones pertaining to the trending AI-generated text detection software (Table 4).

Conclusion and implications

The ability to detect AI-generated essays is getting increasingly important as the use of AI in text generation continues to grow. In this paper, we described a language model to meet the challenge of achieving high accuracy in distinguishing ChatGPT-generated essays from human written essays with a particular focus on not falsely classifying genuinely human-written essays as AI-generated. We developed an n-gram BOWs discrepancy language model which we then used as input to a ML classifier which we trained to predict whether an essay was human or ChatGPT (AI)-generated. We discussed the importance of F2 score for measuring classification performance and totally eliminating False Negatives. We suggest using SVM algorithm upon the proposed model as we recorded 100% accuracy for identifying human generated essays and 92.7% overall accuracy. Although we attained higher overall accuracies (e.g., 94%) via other algorithms and other AI-generated text detection software introduced by OpenAI, GPTZero and Copyleaks yielded superior overall accuracy performances (95.3–96.7%), only SVM provided the perfect Recall with the highest F2 score (97.2%) and no False Negatives ensuring all genuinely human written essays were correctly classified as human generated. Whether a fad or a tech revolution, with its sophisticated AI-led natural language text generation capability, ChatGPT has already started to raise serious concerns over student cheating and as it keeps evolving, it will indisputably keep changing the way we hunt down and defeat academic deceit.

Table 4. Classification performance output of the proposed model algorithms and the trending AI-text detection software

		TPs	TNs	FPs	FNs	Precision	Recall	Acc	F2
Proposed Model	*SVM*	75	64	11	0	0.872	1	0.927	0.972
	EL	70	71	4	5	0.946	0.933	0.940	0.936
	NN	70	70	5	5	0.933	0.933	0.933	0.933
	RF	68	67	8	7	0.895	0.907	0.900	0.904
	LogReg	64	66	9	11	0.877	0.853	0.867	0.858
	NBayes	63	63	12	12	0.840	0.840	0.840	0.840
OpenAI Detector		73	72	3	2	0.961	0.973	0.967	0.971
GPTZero		72	72	3	3	0.960	0.960	0.960	0.960
Copyleaks		70	73	2	5	0.972	0.933	0.953	0.941

Note(s): N = 150.75: Human generated; 75: AI-generated
Source(s): Authors' own creation

It is a challenge for educators to distinguish AI-generated essays from genuinely human (student) generated essays. Therefore, school administrators and leaders must implement AI-generated text detection software to uphold academic integrity standards. AI generated essays submitted by students can be detected by teachers and educational designers using the proposed language model and ML classifier at a high accuracy. Human (student) generated essays can and must be correctly identified with 100% accuracy even if the overall classification accuracy performance is slightly reduced.

Limitations and recommendations

For the proposed n-gram discrepancy language model to function properly on the ML classifier, the essays to be classified must contain sufficient number of words. In this study, we did not include any essay that contained fewer than 400 words because if we did, the classifier would not be able to identify enough n-gram distinctions between human and AI-generated essays. It should be noted that this threshold (400) is arbitrary, hence could be lower or higher. We recommend future research to look into word count as a classification performance indicator.

References

Brown, O. (2023), "The Story of ChatGPT and OpenAI: the Evolution of GPT models", *Medium*, available at: https://medium.com/illumination/the-story-of-chatgpt-and-openai-the-evolution-of-gpt-models-abf201316a9

Copyleaks (2023), "AI-content detector", available at: https://copyleaks.com/features/ai-content-detector (accessed 14 March 2023).

Dwivedi, Y.K., Kshetri, N., Hughes, L., Slade, E.L., Jeyaraj, A., Kar, A.K., *et al.* (2023), "'So what if ChatGPT wrote it?' Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy", *International Journal of Information Management*, Vol. 71, 102642.

Eddy, S.R. (1998), "Profile hidden Markov models", *Bioinformatics (Oxford)*, Vol. 14 No. 9, pp. 755-763.

Fyfe, P. (2022), "How to cheat on your final paper: assigning AI for student writing", *AI and Society*, pp. 1-11, doi: 10.1007/s00146-022-01397-z.

Gatt, A. and Krahmer, E. (2018), "Survey of the state of the art in natural language generation: core tasks, applications and evaluation", *Journal of Artificial Intelligence Research*, Vol. 61, pp. 65-170.

Heikkilä, M. (2022), *How to Spot AI-Generated Text*, MIT Technology Review, Vol. 19 December, available at: https://www.technologyreview.com/2022/12/19/1065596/how-to-spot-ai-generated-text/

Hermans, M. and Schrauwen, B. (2013), "Training and analysing deep recurrent neural networks", *Part of Advances in Neural Information Processing Systems 26 (NIPS 2013)*.

Knill, K. and Young, S. (1997), "Hidden Markov models in speech and language processing", in Young, S. and Bloothooft, G. (Eds), *Corpus-based Methods in Language and Speech Processing*, Kluwer, Dordrecht, pp. 27-68.

Kumar, K. and Thakur, G.S.M. (2012), "Advanced applications of neural networks and artificial intelligence: a review", *International Journal of Information Technology and Computer Science*, Vol. 4 No. 6, p. 57.

Kupiec, J. (1992), "Robust part-of-speech tagging using a hidden Markov model", *Computer Speech & Language*, Vol. 6 No. 3, pp. 225-242.

Love, D. (2014), *No One's Talking about the Amazing Chatbot that Passed the Turing Test 3 Years Ago*, Business Insider, available at: https://www.businessinsider.com/rollo-carpenter-and-cleverbot-2014-6

McKeown, K.R. (1985), "Discourse strategies for generating natural-language text", *Artificial Intelligence*, Vol. 27 No. 1, pp. 1-41.

Olsson, A. and Engelbrektsson, O. (2022), *A Thesis that Writes Itself: On the Threat of AI-Generated Essays within Academia*, working paper, Diva Portal.

OpenAI (2023b), "ChatGPT Jan 9 Version. Prompts: 'Does ChatGPT use neural text generation to generate text?' and 'Expand on this'", available at: https://chat.openai.com/chat (accessed 27 February 2023).

Saul, D. (2023), *ChatGPT Parent Open AI Gets 'Game Changing' Multibillion-Dollar Boost from Microsoft*, Forbes Business, Vol. 23 January, available at: https://www.forbes.com/sites/dereksaul/2023/01/23/chatgpt-parent-open-ai-gets-game-changing-multibillion-dollar-boost-from-microsoft/?sh=57ce64913c0c

Sharples, M. (2022), "Automated essay writing: an AIED opinion", *International Journal of Artificial Intelligence in Education*, Vol. 32 No. 2, pp. 1119-1126.

Turnitin (2020), "How teachers can prepare for AI-based writing", Vol. 21 May, *Turnitin blog*, available at: https://www.turnitin.com/blog/how-teachers-can-prepare-for-ai-based-writing

Weizenbaum, J. (1966), "ELIZA-a computer program for the study of natural language communication between man and machine", *Communications of the ACM*, Vol. 9 No. 1, pp. 36-45.

Winograd, T. (1970), "What does it mean to understand language?", *Cognitive Science*, Vol. 4 No. 3, pp. 209-241.

Yeadon, W., Inyang, O.O., Mizouri, A., Peach, A. and Testrow, C. (2023), "The death of the short-form physics essay in the coming AI revolution", *Physics Education*, Vol. 58 No. 3, 035027.

Further reading

Gptzero (2023), "AI-involvement checker", available at: https://gptzero.me (accessed 15 March 2023).

OpenAI (2023a), "GPT-2 output detector demo", available at: https://openai-openai-detector.hf.space (accessed 16 March 2023).

Corresponding author

Ilker Cingillioglu can be contacted at: ilker.cingillioglu@sydney.edu.au