Date: Sep 25, 2023
To: "Sk Imran Hossain" imran@cse.kuet.ac.bd
cc: "Md. Shahidul Salim" ss@cse.kuet.ac.bd
From: "Scientific Editor" dib-me@elsevier.com
Subject: Your Data in Brief Submission: DIB-D-23-01547

Manuscript No.: DIB-D-23-01547
Title: An Applied Statistics dataset for human vs AI-generated answer classification
Journal Title: Data in Brief
Corresponding Author: Dr. Sk Imran Hossain
All Authors: Md. Shahidul Salim; Sk Imran Hossain
Submit Date: Aug 27, 2023

Dear Dr. Hossain:

Thank you again for your submission to Data in Brief. Your article will require revision before it can be accepted for publication.

I invite you to revise and resubmit your manuscript after having thoughtfully and carefully addressed the comments below and revising your manuscript accordingly.

I look forward to receiving your revised manuscript by
Nov 24, 2023.

PLEASE NOTE: Please submit your revised manuscript before the given due date as a clean file without comments or tracked changes. Please upload a second version with clear highlights by using the 'Track Changes' function in Microsoft Word, so that changes are easily visible to the editors and reviewers. Please provide a letter to editor to explain point by point the details of the revision and the response to the reviewers' comments. Usually authors are only permitted to revise their article twice for Data in Brief, so carefully address all comments, including formatting requests, when revising your manuscript. If you have any questions, please do not hesitate to contact dib-me@elsevier.com.

Yours sincerely,

Scientific Editor
Data in Brief

Reviewers (if applicable):

Reviewer's Responses to Questions

1) Are these data original and produced by the authors?

Please respond with Yes OR No OR N/A.

Reviewer #1: Yes the data is original. As far as I can tell it should be produced by the authors, but I have no way of verifying this.

---

2) Are these data secondary (e.g. censuses, government databases, organizational records)?

Please respond with Yes OR No OR N/A. If YES, please answer 2a, 2b & 2c; if NO go to 3

Reviewer #1: No

2a) Secondary Data Only: were these data collected using variables that make the study unique?

Please respond with Yes OR No OR N/A.

Reviewer #1: N/A

2b) Secondary Data Only: is this collection of secondary data of value to the research community?

Please respond with Yes OR No OR N/A.

Reviewer #1: N/A

2c) Secondary Data Only: do the authors provide the protocol for collecting/creating these data?

Please respond with Yes OR No OR N/A.

Reviewer #1: N/A

---

3) Have the authors used a questionnaire or survey?

Please respond with Yes OR No OR N/A. If YES, please answer 3a; if NO go to 4.

Reviewer #1: No.

3a) Is the sampling representative of the population and rigorously following a scientific method?

Please comment on the rigor of the sampling method and if additional sampling or a different sampling method is required. Please also mention if the questionnaire/survey being used is direct, unambiguous and unbiased.

Reviewer #1: N/A

---

4) Do the authors adequately explain to the research community the utility of these data in the "Value of data" section?

Please include a comment on the validity of this section. Include notes on how this can be improved, if necessary.

Reviewer #1: Yes, the authors did explain the value of the data.

---

5) Are these data described clearly in the "Data description" section?

Please provide suggestions to the author(s) on how to further clarify the presentation and description of the dataset.

Reviewer #1: Yes.

I have the following questions to the authors:
- How did you make sure that the students did solve the question on their own and did not use AI tools like ChatGPT to generate the answers? I saw at the end that you used Turn-it-in for plagiarism check, but I was not sure what the context of "plagiarism" here means? My hunch is that this was used to check if multiple students submitted the same answer. However, what about using ChatGPT to generate the answers? As you stated in the motivation, current detection tools do not work well, so how did you make sure that the students did not use AI to generate the answers?
- Give that the each student randomly answered 50 questions, how did you guarantee that each question had the same amount of answers at the end? You could easily end up with unbalanced data-set, where some questions would have more answers than others? I think it would be good if the tool can balance the random selection so that you get similar or equal answers at the end.

---

6) Is the protocol/method for generating these data adequately described in "Experimental design, materials, and methods" section?

Please include suggestions on how the section can be improved to aid reproducibility/reusability.

Reviewer #1: Yes.

One thing that I noticed which was missing is the consent form. This should be done when the user registers so as to allow the collector to get the users' consent.

---

7) Have the authors provided all the raw data related to all the tables, graphs, images and charts, etc. and are they freely accessible?

Please provide suggestions to the author(s) on how to improve data accessibility for wider usage. Please mention missing raw data, if any.

4

Reviewer #1: The raw data were provided and freely accessible.

---

8) If this data article is related to an existing primary research article is there any duplication? If yes, please comment on this.

Please mention any overlapping text, images, etc.

Reviewer #1: No. The authors did not mention anything related to this.

Reviewer #1: Thank you for submitting your data and short brief to the journal. I have the following questions:

- In the manuscript, the authors mentioned on several occasions "selected by a domain expert", I assume that this means a single person choosing the questions. I believe that this is rather small, and it would have been better to get more experts involved in collecting the questions. Additionally some details are missing, things like: what were the expert qualifications and what they were suited for this task? Based on what did they select the questions? What were the selection criteria?

- Have you obtained an Institutional Review Boards (IRB). This is to ensure that you comply with applicable regulations, meet commonly accepted ethical standards, follow institutional policies, and adequately protect the research participants.

- How did you make sure that the students did solve the question on their own and did not use AI tools like ChatGPT to generate the answers? I saw at the end that you used Turn-it-in for plagiarism check, but I was not sure what the context of "plagiarism" here means? My hunch is that this was used to check if multiple students submitted the same answer. However, what about using ChatGPT to generate the answers? As you stated in the motivation, current detection tools do not work well, so how did you make sure that the students did not use AI to generate the answers?

- Give that the each student randomly answered 50 questions, how did you guarantee that each question had the same amount of answers at the end? You could easily end up with unbalanced data-set, where some questions would have more answers than others? I think it would be good if the tool can balance the random selection so that you get similar or equal answers at the end.

Also there were a couple of small issues and typos here and there, please make sure to go over the text and fix them. Here are two of those that I noted:
- an equal no --> an equal number
- AI detector tools --> AI detection tools

Finally, a suggestion from my side is maybe to add some examples on how such data can be useful by citing some papers that either have used similar data sets within the same contexts. There were some recent papers that came out that had used something similar for both using the data and training AI detection tools:

- Ibrahim, H., Liu, F., Asim, R. et al. Perception, performance, and detectability of conversational artificial intelligence across 32 university courses. Sci Rep 13, 12187 (2023). https://doi.org/10.1038/s41598-023-38964-3
- Vasilatos, Christoforos, Manaar Alam, Talal Rahwan, Yasir Zaki, and Michail Maniatakos. "HowkGPT: Investigating the Detection of ChatGPT-generated University Student Homework through Context-Aware Perplexity Analysis." arXiv preprint arXiv:2305.18226 (2023).

Handling editor -

The paper presents a domain-specific dataset for training AI-generated text detection tools. Particularly, the paper creates a dataset on Applied Statistics with 4,478 question-answer pairs where answers for 116 questions are authored by 100 volunteer students and the equal number of answers are sampled from ChatGPT. Given the emergence of LLM, the paper contributes an interesting and timely dataset. The framework of constructing the presented dataset could be also extended to other domains.

Overall, the paper is well-structured and easy to follow. The present dataset could be useful to the community.

It would be strongly suggested that the paper could provide more details on the dataset as follows:
1. Analysis on the differences between student-authored answers vs. ChatGPT-generated answers in terms of the number of unique words, the average length of answers, the diversity for answers for the same questions, etc.

2. Quality control strategies on collecting answers from volunteer students.

3. Benchmark results of a baseline detection tool trained on the curated dataset.

****

*Note: We cannot accommodate PDF manuscript files for production purposes. We also ask that when submitting your revision you follow the journal formatting guidelines. Figures and tables may be embedded within the source file for the submission as long as they are of sufficient visual quality. For any figure that cannot be embedded within the source file (such as *.PSD, the Photoshop files), the original figure needs to be uploaded separately. Refer to the Guide for Authors for additional information.

To submit your revision, please go to https://www.editorialmanager.com/dib/ and login as an Author.

Your username is: ********

If you need to retrieve your password, please go to:

****

PLEASE NOTE: Data in Brief would like to enrich its relevant online articles with the interactive network diagrams created with Cytoscape. Hence, if applicable, we would like to invite you to upload relevant Cytoscape network files with the revised version of your manuscript to our online submission system. Elsevier will generate the interactive viewer for your datasets and include it with the online article on ScienceDirect. More information can be found at: http://www.elsevier.com/about/content-innovation/cytoscape

This journal uses the Elsevier Article Transfer Service. This means that if an editor feels your manuscript is more suitable for an alternative journal, then you might be asked to consider transferring the manuscript to such a journal. The recommendation might be provided by a Journal Editor, a dedicated Scientific Managing Editor, a tool assisted recommendation, or a combination. For more details see the journal guide for authors.

At Elsevier, we want to help all our authors to stay safe when publishing. Please be aware of fraudulent messages requesting money in return for the publication of your paper. If you are publishing open access with Elsevier, bear in mind that we will never request payment before the paper has been accepted. We have prepared some guidelines (https://www.elsevier.com/connect/authors-update/seven-top-tips-on-stopping-apc-scams) that you may find helpful, including a short video on Identifying fake acceptance letters (https://www.youtube.com/watch?v=o5l8thD9XtE). Please remember that you can contact Elsevier s Researcher Support team (https://service.elsevier.com/app/home/supporthub/publishing/) at any time if you have questions about your manuscript, and you can log into Editorial Manager to check the status of your manuscript (https://service.elsevier.com/app/answers/detail/a_id/29155/c/10530/supporthub/publishing/kw/status/).
