Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Authorship Forensics Portal

Robert SCHMIDT[a], Maiga CHANG[a*], Hsiang-Han CHENG[b], Greg FREDIN[a], Kevin HAGHIGHAT[a] & Rita KUO[c]
[a]*Athabasca University, Canada*
[b]*National Dong Hwa University, Taiwan*
[c]*Utah Valley University, USA*
*maiga.chang@gmail.com

Abstract: This paper presents the research outcome, Authorship Forensics Portal, leveraging both Statistical Natural Language Processing (SNLP) and Convolutional Neural Networks (CNN) techniques to differentiate documents written by humans and ChatGPTs. The portal allows teachers to (1) upload labeled data that contains written text and its author; (2) configure parameters that are required for training models, e.g., 2-class (i.e., human and ChatGPT) or 3-class (i.e., human, ChatGPT 3.5, and ChatGPT as well as the train/test set split ratio, validation set ratio, and validation accuracy threshold for stopping the training process; (3) review the details of a trained model, e.g., the train/test set, the time spent, the prediction results like numbers, true positive, false positive, precision, recall, and f-value, etc.; (4) make their own trained models be private so only themselves can see and use or be public so other teachers can also see and use; and, (5) ask a chosen trained model for its opinion on whether a piece of text written by human or generative AI (e.g., ChatGPT for 2-class prediction and ChatGPT 3.5 or ChatGPT 4 for 3-class prediction). The results demonstrate a significant ability of the models to distinguish between human and AI-written text, with highest precision 0.9868 ($F_{0.5}$ score 0.9647) for the 2-class (human and ChatGPT) testing subset and highest precision 0.9875 ($F_{0.5}$ score 0.9753) for the 3-class (human, ChatGPT 3.5, and ChatGPT 4) testing subset.

Keywords: Natural Language Processing, Statistical NLP, Neural NLP, Convolutional Neural Networks, Part of Speech, ChatGPT.

1. Introduction

Generative Artificial Intelligence (GenAI) (e.g., ChatGPT) is now well-known and popular with the public. Cotton and colleagues (2024) not only point out the potential of students who take advantage to generate content that has higher quality for their assignments, but also could consider as cheating – more like they hire someone (just in this case they hire a bot) to work on the course work for them (Dehouche, 2021); however, teachers have difficulty to distinguish whether a work is written by human or GenAI. OpenAI, the AI research and development company that releases ChatGPT in November 2020, has launched a classifier that can distinguish text written by AI and human authors in January 2023 (Kirchner et al., 2023). However, due to the classifier is not fully reliable – it can identify 26% of AI-written text but misclassify 9% text written by human authors as AI-written ones, which achieves precision 0.74 – OpenAI dismissed the classifier on July 20, 2023.

The research team has done a research with high precision, Authorship Forensics, that adopts statistical natural language processing technique to explore how writers develop distinct language patterns and trains Convolutional Neural Network (CNN) models to differentiate AI-written texts from human-written ones accordingly. The research team develops an open access Authorship Forensics Portal that allows teachers to upload their own data, train models (with half of hours or shorter) for private use or open to the public, access the trained models opened to the public, and ask the chosen trained model to give its opinion on whether or not a given text is written by humans or generative AI.

Section 2 introduces the Authorship Forensics Portal the research team has developed. The preliminary pilot and its results are discussed in Section 3. Moreover, Section 3 also highlights the next step of Authorship Forensics research aiming to clear teachers' doubts and

increase their willingness of consulting the trained models' opinions on whether a text-based work text is written by human student or computer.

2. Authorship Forensics Portal

Authorship Forensics Portal is opened to teachers so they can upload and train models with their own data. After they click "Authorship Forensics" button on the research website[1] they can see the portal as Figure 1 below shows. Teachers can prepare the data according to the format guidelines and configure the parameters they want the models to be trained. The training process is unattended, and teachers don't need to do anything else other than prepare and upload the data. Usually, a model can be processed and trained in a half of hours or shorter – the time spent on training 3-class models is around five minutes.

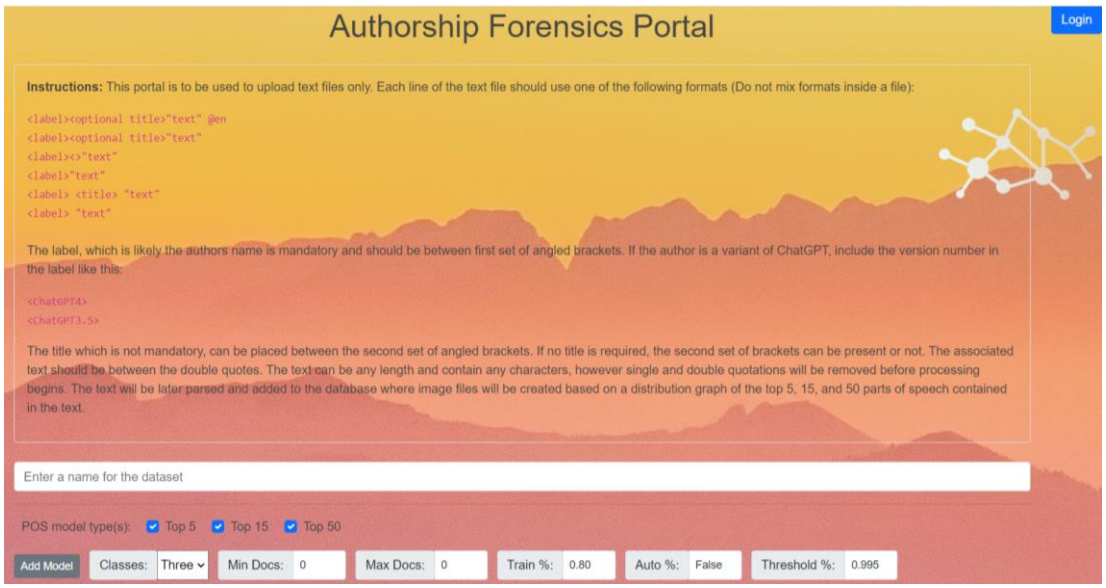

Figure 1. The Authorship Forensics Portal.

Figure 2 shows a list of trained models that are opened to the public can be found at the bottom of the webpage. If a trained model's status is "Finished processing", then teachers can review the details of a trained model by clicking the "View" button. On the other hand, they can also start to use the trained model by clicking its "Predict" button and ask its opinion on whether a given text is written by human student.

Figure 2. Available trained models.

In the detail webpage of a trained model (as Figure 3 shows below) teachers can review the trained model's training history, prediction results on every text, time spent on training a model, and the stats of accuracy, precision, recall, and f-values. They can also copy-and-paste a text, choose a trained model, and ask the chosen model's opinion on whether the given text is written by human or ChatGPT (or ChatGPT 3.5 or 4) as Figure 4 shows below.

[1] https://ngrampos.vipresearch.ca/

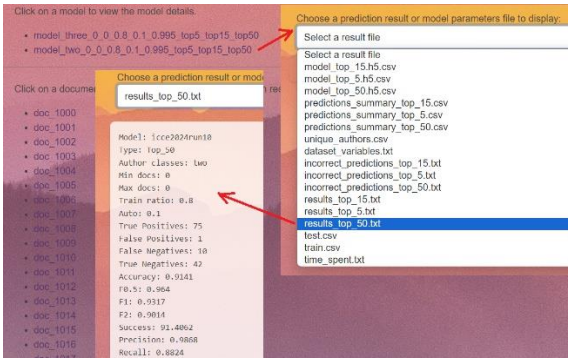

Figure 3. Details of a trained model.

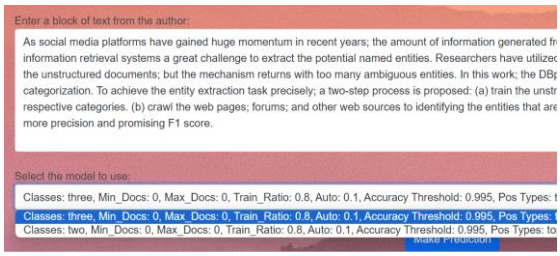

Figure 4. Ask a trained model's opinion.

3. Discussion and Future Work

The preliminary pilot used the data that consists of 212 human-written abstracts of academic papers published by VIP Research Group[2], alongside 424 AI-generated abstracts, evenly divided between ChatGPT versions 3.5 and 4. The common top 50 POS tags were considering as the written pattern features. The research created two datasets, 2-class has all text labelled with human and ChatGPT and 3-class dataset has data labelled with human, ChatGPT 3.5, and ChatGPT 4. Both datasets were duplicated, and the text was repeatedly and randomly assigned to training and testing subsets based on 80:20 ratios. At the end the pilot trained the models 10 times on the training subsets (in which 10% of text were split for validation purpose) until their validation accuracy exceeded 99.5% (to avoid overfitting) and the trained models' performances were verified with the testing subsets accordingly.

This research aims to provide teachers advice on the possibility of a course work submitted by their students were written by ChatGPT. In scenarios where ethical considerations are paramount, the emphasis on the precision metric becomes crucial in avoiding false accusations against human authors. The results demonstrate that the trained models are capable of distinguishing human and AI-written text, with average/highest precision 0.9323/0.9868 (where $F_{0.5}$ score 0.9356/0.9647) for the 2-classt and average/highest precision 0.9445/0.9875 (where $F_{0.5}$ score 0.9449/0.9753) for the 3-class. The results outperform not only the precision 0.74 that OpenAI's AI classifier has, but also the precision 0.8929 ($F_{0.5}$ score 0.9124) and precision 0.9259 ($F_{0.5}$ score 0.9398) at document level that the XGBoost classifier has for differentiating texts written by ChatGPT 3.5 and ChatGPT 4 (Desaire et al., 2023).

While the trained model's performances have outperformed almost all existing research, the research team is still planning to conduct more evaluations on the reusability of trained models and the comparisons with existing online detectors to clear teachers doubts on using the Authorship Forensics and make them have faith on the GenAI-written detection results.

References

Cotton, D. R. E., Cotton, P. A., Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 61(2), 228-239. https://doi.org/10.1080/14703297.2023.2190148

Dehouche, N. (2021). Plagiarism in the age of massive generative pre-trained transformers (GPT-3). Ethics in Science and Environmental Politics, 2, 17-23. https://doi.org/10.3354/esep00195

Desaire, H., Chua, A. E., Kim, M.-G., & Hua, D. (2023). Accurately detecting AI text when ChatGPT is told to write like a chemist. Cell Reports Physical Science 4, 101672. https://doi.org/10.1016/j.xcrp.2023.101672

Kirchner, J. H., Ahmad, L., Aaronson, S., & Leike, J. (2023). New AI classifier for indicating AI-written text. OpenAI. Access: https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/

[2] the journal and conference papers presented at https://maiga.athabascau.ca/