CS 309 Ethics Project
AI-based detectors and its core technologies

Abdon Morales
The University of Texas at Austin
Department of Computer Science

December 7, 2024

Chapter I

Technology Description

With the rise of large language models, such as ChatGPT and Claude, the response that the majority of the higher and K-12 education system(s) across the United States and beyond have implemented is the use of AI-based or AI-integrated detection systems to prevent the use of AI for academic dishonesty. Major sites and companies like Turnitin, ZeroGPT, Stanford's MOSS, and other companies have created these detection systems to prevent academic dishonesty using a confidence or probability score to the user, the professor.

What is AI-based detector? An AI-based detector is more of a multimodal detection system infused with AI; it can be tailored for code syntax analysis, checking for similarities in papers, grammar-check on an email, looking at vocabulary usage.

Large language models in higher education and K-12, has been considered both a positive impact in education, such as giving a helping hand to the instructor on teaching students concepts they still don't understand via LLMs [chatbots] when the instructor is not available. However, it can have a negative impact on education, such as academic dishonesty and not grasping the concepts fully. This could also lead to a higher yield in procrastination[1], which becomes yet another negative factor in one's educational endeavors.

This technology interests me because it addresses the ethical challenges in education and content verification while displaying artificial intelligence's practical applications in maintaining academic integrity.

[1]Muhammad Abbas, Farooq Ahmed Jam, and Tariq Iqbal Khan (2024). "Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students". In: *International Journal of Educational Technology in Higher Education* 21.1, p. 10. DOI: 10.1186/s41239-024-00444-7. URL: https://doi.org/10.1186/s41239-024-00444-7.

Chapter II

What makes this an AI-based technology?

What is artificial intelligence? We know that it is a science and a set of computational technologies that are inspired by, but typically operate quite differently from; the ways people use their nervous systems and bodies to sense, learn, reason, and take action. In other words, it's a compilations of technologies ranging from reinforced learning, natural language processing, planning, symbolic reasoning, ..., and human computation.

In addition, many detection systems today rely on AI paradigms that aligns with the earlier definition of artificial intelligence. Natural language processing (NLP) analyzes and classifies textual data, such as identifying factual information; while symbolic reasoning applies rules to draw inferences and reinforcement learning improves performance through feedback loops. These paradigms combine diverse techniques, inspired by human cognition, to address complex problems efficiently.

Before the introduction of artificial intelligence technologies in the areas of education, journalism, and now fact-checking; the jobs of these detectors used to be more human-centric such as a professor, moderator, or a senior editorial of some paper. However, around 2015, many of these human detectors were replaced with modern AI/ML-based detectors.

These AI-based detectors are largely dependent on the companies' usage, while the majority of these systems use machine learning and large datasets. In addition, much of the core technologies of these AI-based detectors are based on the artificial intelligence paradigms of a Symbolic Rule-based Approach [Machine Learning], and neural networks due to its ability to improve and learn.

Chapter III

The strengths and limitations

Some strengths of an AI-based detector is improving text inputs such as essays, a sentence, vocabulary choice [word choice], detecting bugs in a code review, and even threat detection in the cybersecurity space. These are some of the technological advantages that artificial intelligence has a slight edge over its human counterpart, however, there are some improvements that can be introduce to make these AI-based detection systems to become more accurate such as threat detection where it has to be 100% correct and accurate, such as the conceptual system[1] induced by Dr. Rahul Mishra in a 2023 IEEE paper. Yet, these systems have their limitations that their human counterparts can excel.

Some of the limitations that AI-based detection systems that have become persistent has been their accuracy in the areas of academic dishonesty and AI similarity. For example, if I were to type up an essay and then generate another essay using something like ChatGPT or Claude; then run it through an AI detection system like Turnitin or ZeroGPT, the chances of it detecting the AI-generated essay is 85-100%. However, according to Lauren Coffey and Turnitin, Coffey[2] emphasize "Turnitin says its AI-detection tool, in an attempt to avoid false positives, can miss roughly 15 percent of AI-generated text in a document..." while Turnitin[3] claims in their guide for AI writing, "In order to maintain this low rate of 1% for false positives,

[1] Rahul Mishra (2023). "Cyber Security Threat Detection Model Using Artificial Intelligence Technology". In: *2023 2nd International Conference on Edge Computing and Applications (ICECAA)*, pp. 324–330. DOI: 10.1109/ICECAA58104.2023.10212209.

[2] Lauren Coffey (2024). *Professors Cautious of Tools to Detect AI-Generated Writing*. URL: https://www.insidehighered.com/news/tech-innovation/artificial-intelligence/2024/02/09/professors-proceed-caution-using-ai.

[3] Turnitin (2024). *How does Turnitin ensure that the false positive rate for a document remains less than 1%?* Webpage. Accessed: 2024-11-25. URL: https://guides.turnitin.com/hc/en-us/articles/28477544839821-Turnitin-s-AI-writing-detection-capabilities-FAQs#h_01J2HRTB141222MRDXMW24VVJC.

there is a chance that we might miss 15% of AI written text in a document. We're comfortable with that since we do not want to incorrectly highlight human written text as AI-written." This leaves us with the question, what happens to those 15% of undergraduates that are incorrectly identified by the detection system?

Another limitations seen in these systems, is the issue of credibility versus accuracy. This is an issue that is problematic in the majority of these AI-based detection systems. This can be seen in journalism as the majority of major news corporations have turned to automated fact-checking systems which also have their severe limitations in the case of credibility versus accuracy. Furthermore, systems that infer credibility based on source reliability risk overlooking individual inaccuracies from trusted sources as per Lucas Graves' Factsheet[4] on understanding the limits of AFC [Automated Fact-Checking], "The most dangerous misinformation for each of us comes from the sources we trust".

However, some strengths that the Automated Fact-Checking system has is a strong identification system for check-worthy claims, flagging of repeated false information, and encouragement of open data standards. With a strong identification systems for check-worth claims of newspapers, articles, blogs, and other medias; this makes it efficient for identifying for what's false information, and validated and factual information for the human fact-checkers [supervisors]. Furthermore, the encouragement of adopting open data standards makes it easier for developers to create a more diverse and accurate systems without sacrificing credibility as mentioned by Dr. Graves, "Both Chequeado and Full Fact...have campaigned to make official statistics available as structured data friendly for developers"[5].

[4]Lucas Graves (2018). *Understanding the Promise and Limits of Automated Fact-Checking*. Factsheet. DOI: 10.60625/risj-nqnx-bg89. Reuters Institute for the Study of Journalism. URL: https://reutersinstitute.politics.ox.ac.uk/our-research/understanding-promise-and-limits-automated-fact-checking.

[5]Graves 2018.

Chapter IV

Addressing the risks

Some of risks associated with AI-based detection systems that have become a persistent issue is the following:

1. Privacy and data concerns of the end-user.
2. Lack of transparency in the decision-making process in its algorithm.
3. Potential bias(es) in the data that the model [system] is trained.

Privacy concerns have become a spotlight issue for the United States and private companies determining how secure one's data is when using these systems. In the area of education the concerns of storing student data and entering into these detectors that are managed by a some company like Turnitin; this may also have a conflict of interest with major educational institutions[1] in terms of their own privacy and data management policies.

Another risk that has become heavily associated with these AI-based detection systems is the lack of transparency in the decision-making process in the machine learning models of these systems. This lack of transparency comes from the complex system of algorithms that, in exchange, makes it difficult to understand how decisions are made. In high-stakes situations, this can be catastrophic as people affected by these systems may have no way to know or challenge the results.

Bias in both the training and data sets is a persistent issue in AI-based technologies, including large-language models like ChatGPT;

[1] Vanderbilt University (2023). *Guidance on AI Detection and Why We're Disabling Turnitin's AI Detector*. Accessed: 2024-12-02. URL: https://www.vanderbilt.edu/brightspace/2023/08/16/guidance-on-ai-detection-and-why-were-disabling-turnitins-ai-detector/.

and detection systems such as fact-checking in journalism, and potential racial bias[2] in cheating detection systems for writing such as Turnitin. These biases pose a great risk to equalized outcomes, as they can reinforce stereotypes, target specific demographics, and lead to discriminatory practices[3]; it also has far-reaching implications towards society, but it undermines the trust of AI systems while exacerbating existing inequalities.

Some practical solutions to address these risks is the implementation of an option to create a local self-hosted instance of the service [system] which reduce the conflict of interest and the data management implications between the institutional end-user's privacy policy, and that of the company who is providing the service. By ensuring privacy and autonomy, this approach aligns with a consequentialist framework by minimizing potential harm to the end-users as well as maximizing the benefits of protecting sensitive information. In addition, we can also attempt to transition [and create] a more open system model to reduce the lack of transparency of the system's algorithms and complexity; for developers, companies, small businesses, education, and government. With this open system model, we can also tackles the issue of bias by creating a more unique training dataset that minimize the ratio of non-biased to biased data. With all of these suggested practical solutions, grounded in a consequentialist approach with the aim to maximize positive outcome, reduce harm, and ensure a much safer and ethical AI-based detection; which will also positively benefit other AI core technologies.

[2] Julia Angwin et al. (2016). *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks.* Accessed: 2024-12-02. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] Jackie Davalos and Leon Yin (2024). "Do AI Detectors Work? Students Face False Cheating Accusations". English. In: *Bloomberg*. Accessed: 2024-12-02. URL: https://www.bloomberg.com/news/features/2024-10-18/do-ai-detectors-work-students-face-false-cheating-accusations.

Chapter V

Improvements

Some improvements that I suggest or recommend for these AI-detection companies could implement in their models, quality assurance, datasets, and algorithms that could both help students, journalists, professors, and other professions are the following:

1. Conduct internal and public beta testing to ensure model accuracy; public input improves triaging processes and diversifies datasets.

2. Minimize the amount of model bias to keep high accuracy, aligned by consequentialist or utilitarian ethics.

3. Open-source the model, datasets, weights, and statistics of the system(s) for greater transparency, the ability for people to contribute to the datasets as way of diversifying the data, and ease of use for developers that are looking to use the model for their own needs or development. In addition, it provides research opportunities for researchers to explore more of the impact of these systems.

These improvements enhance transparency in machine learning, reinforcement learning, bias control, and ethical practices for developers and end-users. An open model fosters collaboration in the open-source community and research opportunities; aligning with a utilitarian approach to inclusivity. For instance, reinforcement learning could be enhanced by incorporating dynamic feedback loops that adapt to the changing user input, further reducing bias over time. Similarly, symbolic reasoning could improve explainability by offering logical justifications for its decisions.

By implement these changes, AI detection systems can evolve to become more robust, equitable, and scalable [than ever before!], meeting the diverse needs of their users while adapting to future challenges.

Chapter VI

Conclusion and future directions

AI-based detection systems, such as Turnitin, ZeroGPT and others, represent a significant advancement in maintaining academic integrity and verifying content through technologies like machine learning, reinforcement learning, and natural language processing. These systems outperform in areas such as identifying AI-generated text, improving text quality, and supporting human oversight. However, they face persistent challenges, including biases in training data, limited to no transparency, and potential privacy violations; which hinders a broader adoption and trust.

To address the challenges, future advancements should focus on open-sourcing models and datasets to foster an environment of transparency and collaboration while reducing bias through unique data contributions. In addition, adopting cutting-edge techniques, such as federated learning, could protect user data while enabling collaborative improvements across institutions. Enhancing reinforcement learning with dynamic feedback loops and employing symbolic reasoning to improve decision-making explainability, are critical steps toward making these systems more user-friendly.

All in all, these improvements not only position AI-based detection systems to meet the diverse needs of educators, students, journalists, consumers, and researchers but also ensure their adaptability to future challenges. Ethically, the balance between automation and human oversight will be essential in preventing harm and promoting fairness. By addressing these limitations and embracing the innovative solutions of the future, AI-based detection systems can continue to evolve into robust, inclusive, and impactful tools for diverse applications.

Bibliography

Abbas, Muhammad, Farooq Ahmed Jam, and Tariq Iqbal Khan (2024). "Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students". In: *International Journal of Educational Technology in Higher Education* 21.1, p. 10. DOI: 10.1186/s41239-024-00444-7. URL: https://doi.org/10.1186/s41239-024-00444-7.

Angwin, Julia et al. (2016). *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks.* Accessed: 2024-12-02. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Coffey, Lauren (2024). *Professors Cautious of Tools to Detect AI-Generated Writing*. URL: https://www.insidehighered.com/news/tech-innovation/artificial-intelligence/2024/02/09/professors-proceed-caution-using-ai.

Davalos, Jackie and Leon Yin (2024). "Do AI Detectors Work? Students Face False Cheating Accusations". English. In: *Bloomberg*. Accessed: 2024-12-02. URL: https://www.bloomberg.com/news/features/2024-10-18/do-ai-detectors-work-students-face-false-cheating-accusations.

Graves, Lucas (2018). *Understanding the Promise and Limits of Automated Fact-Checking*. Factsheet. DOI: 10.60625/risj-nqnx-bg89. Reuters Institute for the Study of Journalism. URL: https://reutersinstitute.politics.ox.ac.uk/our-research/understanding-promise-and-limits-automated-fact-checking.

Mishra, Rahul (2023). "Cyber Security Threat Detection Model Using Artificial Intelligence Technology". In: *2023 2nd International Conference on Edge Computing and Applications (ICECAA)*, pp. 324–330. DOI: 10.1109/ICECAA58104.2023.10212209.

Turnitin (2024). *How does Turnitin ensure that the false positive rate for a document remains less than 1%?* Webpage. Accessed: 2024-11-25. URL: https://guides.turnitin.com/hc/en-us/articles/28477544839821-Turnitin-s-AI-writing-detection-capabilities-FAQs#h_01J2HRTB141222MRDXMW24VVJC.

Vanderbilt University (2023). *Guidance on AI Detection and Why We're Disabling Turnitin's AI Detector*. Accessed: 2024-12-02. URL: https://www.vanderbilt.edu/brightspace/2023/08/16/guidance-on-ai-detection-and-why-were-disabling-turnitins-ai-detector/.