Building a better AI detector using retrieval methods

Dev Bhatia
Computer Systems Lab Dr. Yilmaz
Period: 3
Date: 12/12/2024
Project Presentation 2

Problem

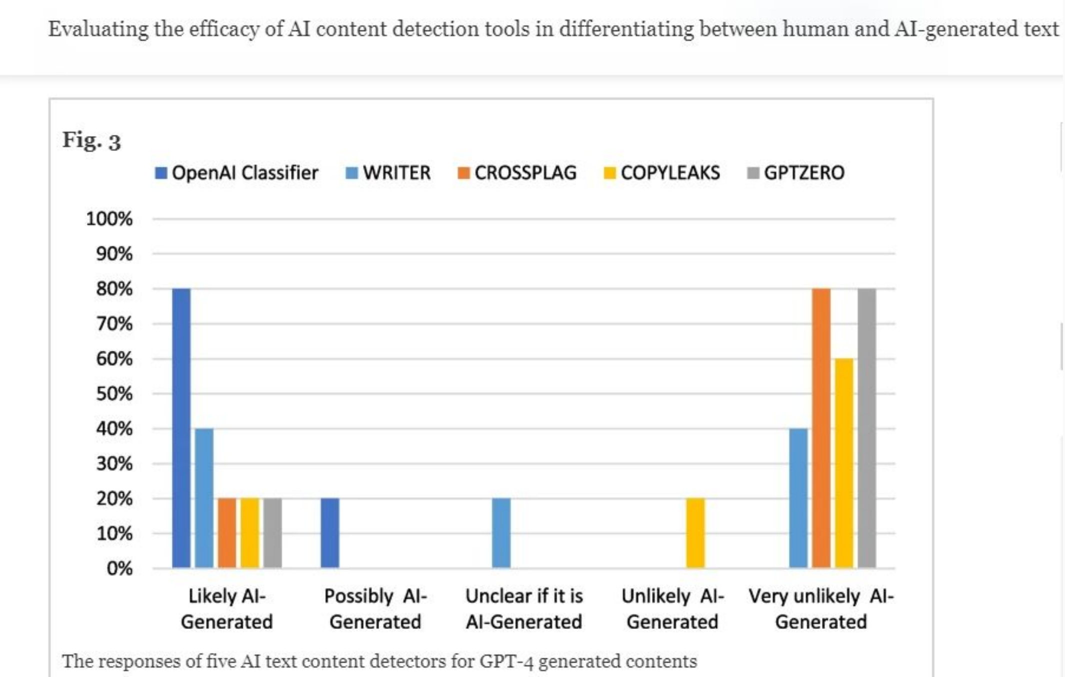
- AI used very prominently
- People tend to paraphrase AI content
- Difficulty telling original work from AI generated text
- Current AI detectors don’t work well
- Leads to lack of knowledge for students
- AI gives untrustworthy or false information
- Need better detection mechanism

Background

- Many AI detectors out there
- Accuracy varies significantly
- Unreliable for teacher use
- Possibility of false positives
- Not close to 100% accuracy

Other Solutions

- Most prominent AI detectors:
    - GPTZero
    - Copyleaks
    - CrossPlag
- Detectors do not perform well against GPT 4.0
- Many false negatives and uncertain classifications



Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text

Fig. 3

The responses of five AI text content detectors for GPT-4 generated contents

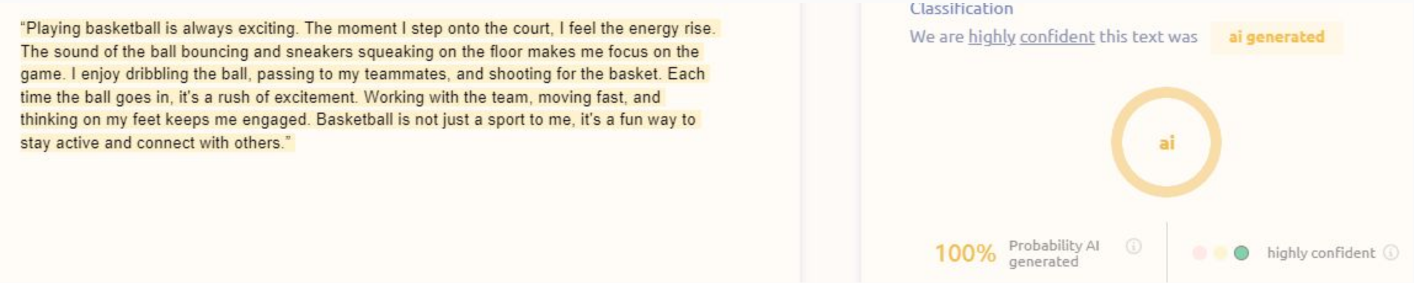
Other Solutions Continued

- How these AI detectors work
    - GPTZero:
        - Amount of predictability in the text
        - Looks at the variance in the sentences, AI generated text typically has less variety
    - Copyleaks
        - Scans the document against different sources to see if it matches anything
        - Sentence by sentence detection against human writing
    - CrossPlag
        - Text analysis using Natural Language Processing
        - Using a dataset created by human and AI content
- Pros: Detect well on older GPT generations
- Cons: Fail on GPT 4.0 and paraphrased AI generations, inconsistencies

Demo of current solutions

- GPTZero
- CopyLeaks: https://copyleaks.com/ (doesn't seem to work without subscription)
- CrossPlag: https://app.crossplag.com/individual/detector
- Sample AI generated text: “Playing basketball is always exciting. The moment I step onto the court, I feel the energy rise. The sound of the ball bouncing and sneakers squeaking on the floor makes me focus on the game. I enjoy dribbling the ball, passing to my teammates, and shooting for the basket. Each time the ball goes in, it’s a rush of excitement. Working with the team, moving fast, and thinking on my feet keeps me engaged. Basketball is not just a sport to me, it’s a fun way to stay active and connect with others.”

AI Detectors with paraphrasing continued



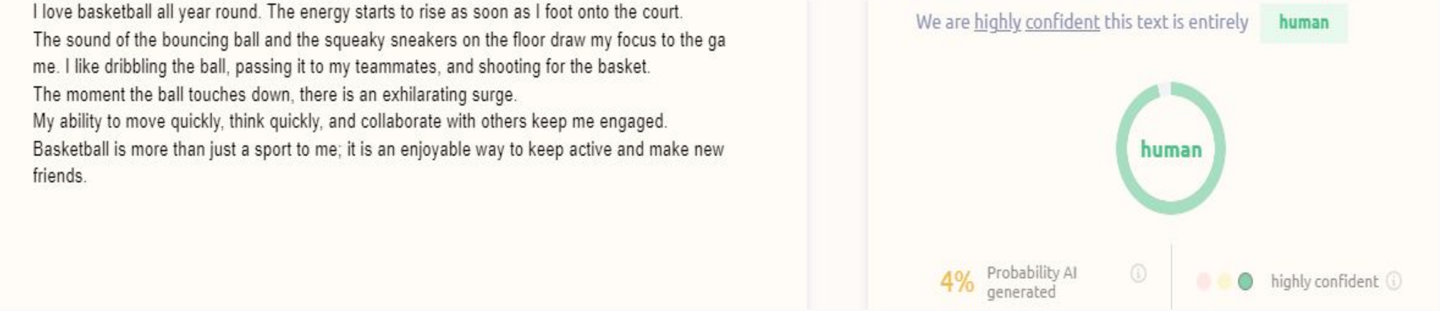
"Playing basketball is always exciting. The moment I step onto the court, I feel the energy rise. The sound of the ball bouncing and sneakers squeaking on the floor makes me focus on the game. I enjoy dribbling the ball, passing to my teammates, and shooting for the basket. Each time the ball goes in, it's a rush of excitement. Working with the team, moving fast, and thinking on my feet keeps me engaged. Basketball is not just a sport to me, it's a fun way to stay active and connect with others."

Classification

We are highly confident this text was ai generated

100% Probability AI generated

highly confident



I love basketball all year round. The energy starts to rise as soon as I foot onto the court.
The sound of the bouncing ball and the squeaky sneakers on the floor draw my focus to the ga
me. I like dribbling the ball, passing it to my teammates, and shooting for the basket.
The moment the ball touches down, there is an exhilarating surge.
My ability to move quickly, think quickly, and collaborate with others keep me engaged.
Basketball is more than just a sport to me; it is an enjoyable way to keep active and make new friends.

We are highly confident this text is entirely human

4% Probability AI generated

highly confident

Why Is Mine Better?

- Model that is more accurate on GPT 4.0
  - Current detectors struggle on this
- Model that is better with paraphrased GPT 4.0
  - GPT 4.0 content that people paraphrase but the meaning is the same
- Model that gives more consistency in detecting GPT 4.0 content

Novelty

- Using retrieval methods (not used on common AI detectors) on GPT 4.0
    - Using a database of AI generations to tell whether text is a paraphrasing of AI generated text
        - Use cosine similarity scores
        - Find any matches to previous generations
    - Accuracy will remain high even as the amount of paraphrasing goes up
- Use on “mixing attacks”
    - Texts that has both human written elements and AI generated text
- Higher accuracy than current AI detecting algorithms have

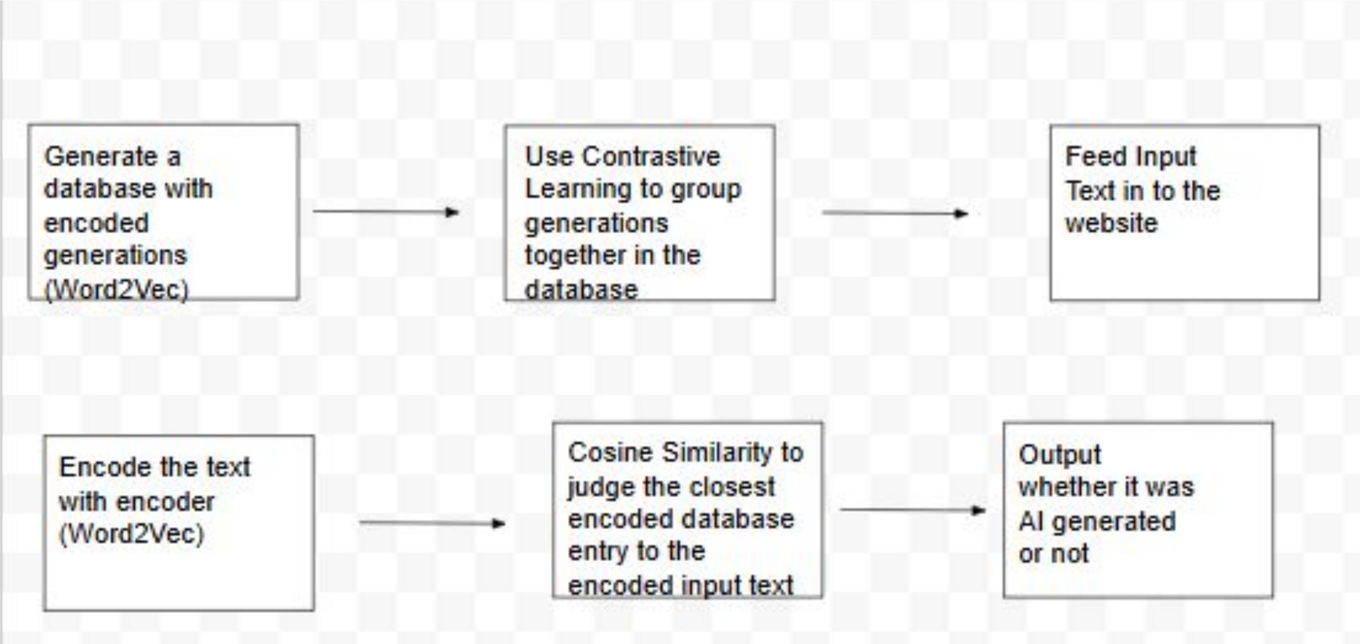
Impact

- Used in school by teachers
  - A more reliable way to check AI content
- Trustworthy source of AI detection
- Maintains integrity
- Creates a fair learning environment

Method

- Store Word2Vec encoded sequences of text that were generated by GPT 4.0 in database in a web server
- Feed in input text
- Use a Word2Vec encoder to encode input text
- Check to see if input text matches database text (cosine similarity)
- Set the threshold (T), if the text score is higher than T, then it is judged as not similar
- Output whether the text is detected as GPT 4.0 generated or not

Method: Systems Architecture



Generate a database with encoded generations (Word2Vec)

Use Contrastive Learning to group generations together in the database

Feed Input Text in to the website

Encode the text with encoder (Word2Vec)

Cosine Similarity to judge the closest encoded database entry to the encoded input text

Output whether it was AI generated or not

Encoding

- Encoding using a Word2Vec Model
    - Finds the semantic relationship between words
    - Captures the meaning of the word
    - 2 Layer Neural Network
    - Continuous Bag of Words
- Getting a vector representation of words
- Uses large corpus of text

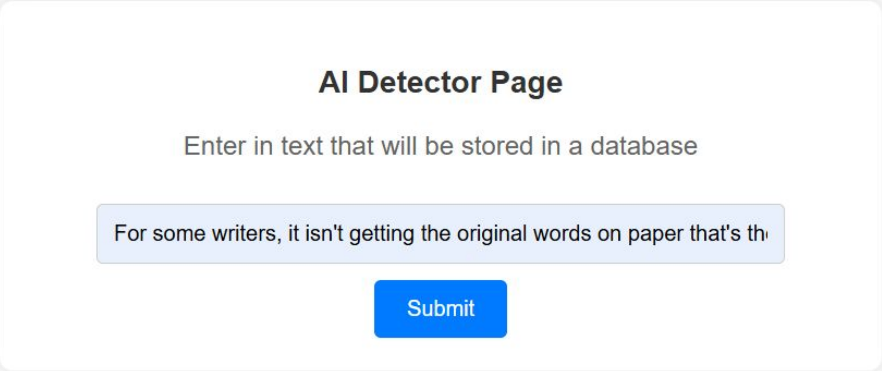
Database

- Created using generations around a prompt or topic
- Teacher can put a prompt in and generate AI responses
- Can generate hundreds of pieces of text for the database
- Take the LLM output and encode it as a vector
- Store the vector in the database (Word2Vec)

Contrastive Learning

- Positive and Negative pairs
  - Positive pairs are associated closer together
    - Paraphrased AI generations and AI output
  - Negative pairs are further
    - Human and machine written text
  - Distance can be used to see similarities between text
- Take closest entry
- Compare input to closest piece of text in the database with cosine similarity
- $$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\|\vec{b}\|}$$
  $$\|\vec{a}\| = \sqrt{a_1^2 + a_2^2 + a_3^2 + \cdots + a_n^2}$$
  $$\|\vec{b}\| = \sqrt{b_1^2 + b_2^2 + b_3^2 + \cdots + b_n^2}$$

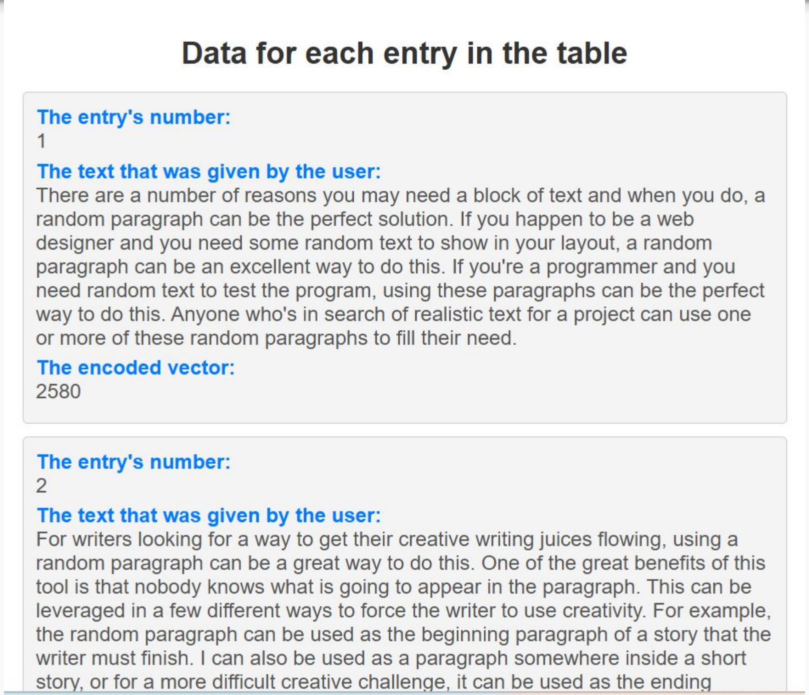
Demo



AI Detector Page

Enter in text that will be stored in a database

For some writers, it isn't getting the original words on paper that's th



Data for each entry in the table

The entry's number:
1

The text that was given by the user:
There are a number of reasons you may need a block of text and when you do, a random paragraph can be the perfect solution. If you happen to be a web designer and you need some random text to show in your layout, a random paragraph can be an excellent way to do this. If you're a programmer and you need random text to test the program, using these paragraphs can be the perfect way to do this. Anyone who's in search of realistic text for a project can use one or more of these random paragraphs to fill their need.

The encoded vector:
2580

The entry's number:
2

The text that was given by the user:
For writers looking for a way to get their creative writing juices flowing, using a random paragraph can be a great way to do this. One of the great benefits of this tool is that nobody knows what is going to appear in the paragraph. This can be leveraged in a few different ways to force the writer to use creativity. For example, the random paragraph can be used as the beginning paragraph of a story that the writer must finish. I can also be used as a paragraph somewhere inside a short story, or for a more difficult creative challenge, it can be used as the ending

Results

- I hope to measure 85% accuracy or better on detecting text generated by GPT 4.0
- Model takes input text
    - Could be paraphrased
    - Could be mixed in with human content,
    - Model tells whether it was AI generated or not
- Expect better performance than current AI detectors

Limitations

- Text not generated by students
- Text with shorter number of words
- Not using watermarking and statistical outlier methods
- Other potential AI’s (besides GPT 4.0)
- Scalability to more than 15 million generations

Future Work

- Improvements in text generated for other things besides school
- Use retrieval method along with other methods (watermarking, statistical outlier) to improve shorter text identification
- Experiment retrieval methods with other AI’s
- Test accuracy with more than 15 million generations

Conclusion

- I solved the problem of AI Detection using Retrieval Methods

References

[1] A. Singh, "A Comparison Study on AI Language Detector," *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA, 2023, pp. 0489-0493, doi: 10.1109/CCWC57344.2023.10099219.

[2] D. Dukić, D. Keča and D. Stipić, "Are You Human? Detecting Bots on Twitter Using BERT," *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, Sydney, NSW, Australia, 2020, pp. 631-636, doi: 10.1109/DSAA49011.2020.00089.

Elkhatat, A.M., Elsaid, K. & Almeer, S. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *Int J Educ Integr* 19, 17 (2023). https://doi.org/10.1007/s40979-023-00140-5

Krishna, Kalpesh, et al. "Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense." *Advances in Neural Information Processing Systems* 36 (2024).

Perkins, Mike, et al. "Simple Techniques to Bypass GenAI Text Detectors: Implications for Inclusive Education: Revista De Universidad y Sociedad Del Conocimiento." *International Journal of Educational Technology in Higher Education*, vol. 21, no. 1, 2024, pp. 53*. ProQuest*, https://www.proquest.com/scholarly-journals/simple-techniques-bypass-genai-text-detectors/docview/3101842024/se-2, doi:https://doi.org/10.1186/s41239-024-00487-w.

[6] GPTZero: https://gptzero.me/

Q&A

THANKS!