Exploring the Cause of Bias Towards Non-native Writers in GPT-Detector

{CSC3160} {Default} Project

Bao Zongbo (120090442)
School of Data Science
Chinese University of Hong Kong, Shenzhen
120090442@link.cuhk.edu.cn

Abstract

A recent research from Stanford indicated that there is bias towards non-native English writers in almost all existing GPT-Detector. It is important to figure out the cause of such bias.In this project we made and validated two hypothesises about the cause of bias. In the end, we proposed some possible way the eliminate such bias.

1 Key Information to include

- I would like to give up 1% in poster session and move the reminding 4% to project report.

2 Introduction

The rapid adoption of generative large language models has brought about substantial advancements in digital communication, while simultaneously raising concerns regarding the potential misuse of AI-generated content.
Although numerous detection model have been proposed to differentiate between AI and human-generated content, a recent research paper [3] from Stanford suggested that those detectors may have bias towards non-native English writer. In another word, non-native writing samples are consistently misclassified as AI-generated, whereas native writing samples are accurately identified.

In order to eliminate such bias from detector, we need to understand how it is unintentionally generated in the training process.
We made two hypothesises for the cause of bias:

- The mismatch of dataset distribution leads to the bias of detector.
- The limited linguistic expressions in non-native writing leads to the bias of detector.

We tested the first hypothesis by fine-tuning the baseline model on two different essay dataset provided in [3]. One dataset was collected from US 8-th Grade Essay, which can be considered as good source of native writing samples. While another dataset was collected from non-native-authored TOEFL essays. The details of Experiment are described in 5. The result shows adding non-native-authored samples can effectively decrease the bias of detector.
We tested the second hypothesis by using the GLTR tools developed in [1] to test the essay written by non-native speaker and native speaker, and the result shows that non-native-authored essay tend to use more normal and repeated phrases and expressions. Additionally, in [3], researchers tried using ChatGPT to improve the word choices by prompts like "Make my linguistic expressions more complex.". We tested our fine-tuned model on their original-polished essay pairs in 5. The result validated the second hypothesis.

3 Related Work

3.1 Generated text detector

To my knowledge, existing detectors can be divided into two types.
First type is fine-tuning the pre-trained languege model(which has the ability to extract features of text) on the dataset that contains human-written and AI-generated samples. One typical example is GPT-Detector proposed in [2], which is a pre-trained Roberta model[4] fine-tuned on a dataset containing the answers from human and ChatGPT to the same question. We use this model as our baseline.
Second type is using perplexity and burstiness of the text as a feature in inference. GPT Zero[5] is a typical example of this. Perplexity is a measurement of how well a language model like ChatGPT can predict the following word based on the previous ones. A low perplexity score means that the language model is confident at its predictions. Burstiness, is used to model the variation of perplexity through the whole passage. A low burstiness score means the perplexity score is stable.

3.2 Bias of Detector

The researchers from Stanford evaluated several commonly-used GPT-detector on native and non-native writing samples in [3]. Take the detector from Open AI as a example, 59% of non-native-authored essay is misclassified as AI-generated, while the number for native-authored essay is merely 8%.

4 Approach

As discussed in 2, the two hypothesises we want to test is:

- The mismatch of dataset distribution leads to the bias of detector.
- The limited linguistic expressions in non-native writing leads to the bias of detector.

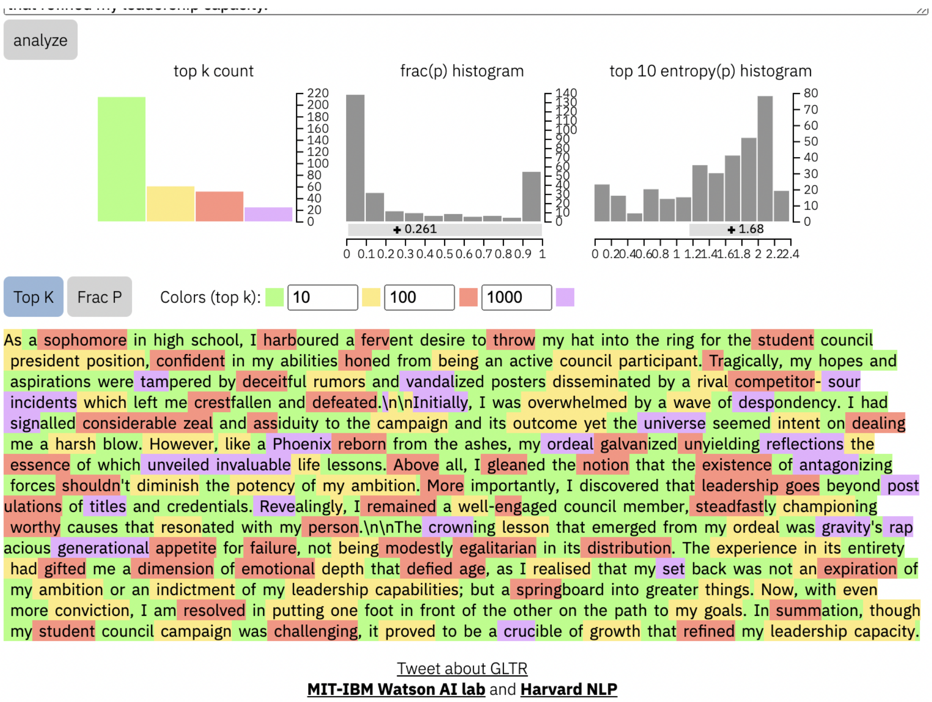
We fine-tuned the pre-trained Roberta model[4] on a dataset[2] containing over 10000 pairs of answers from human and ChatGPT to the same question, and used this model as our baseline. We denoted it as *M_base*.

4.1 First Hypothesis

To test the first hypothesis, we further trained *M_base* on non-native-authored TOEFL essays and corresponding answers from ChatGPT by the same writing prompt. We denoted the trained model as *M_1*. In order to better compare the result, we also trained *M_base* on same amount of native-authored US 8th grade essays and corresponding ChatGPT answers. We denoted the trained model as *M_2*. We evaluated the three models, *M_base*, *M_1* and *M_2* on the test set composed of TOEFL essays and US 8th grade essays, the result are shown in 5.4.

4.2 Second Hypothesis

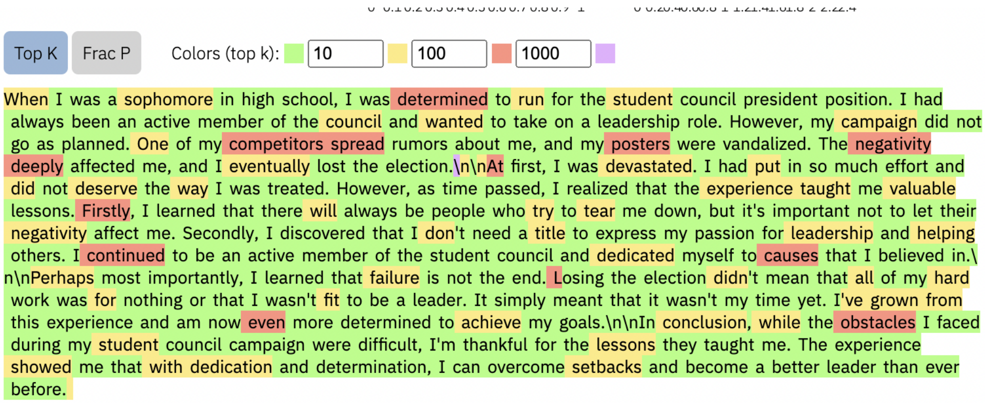
To test the second hypothesis, we feed a pair of essay sample from the dataset provided in [3] to the GLTR tool developed in[1]. GLTR computes the probability of being next word for every word in the vocabulary, and with those probabilities they can get the rank of true word in the whole vocabulary.
The result of GLTR is shown in 1 and 2. As we can see, native writing samples contain more word choices that ranked over 100 than the non-native-authored text. In another word, language model is less confident when making predictions on native-authored text. This result indicated that native-authored text may have higher perplexity for its rich word choices.
However, the influence of word choices richness on the detecting result is still unclear. Therefore, we used ChatGPT to improve the word choices of original TOEFL essays by prompts like "Make my words and expressions more flexible". We tested our models on the obtained dataset, and the result is shown in 5.4.



As a sophomore in high school, I harboured a fervent desire to throw my hat into the ring for the student council president position, confident in my abilities honed from being an active council participant. Tragically, my hopes and aspirations were tampered by deceitful rumors and vandalized posters disseminated by a rival competitor- sour incidents which left me crestfallen and defeated.\n\nInitially, I was overwhelmed by a wave of despondency. I had signalled considerable zeal and assiduity to the campaign and its outcome yet the universe seemed intent on dealing me a harsh blow. However, like a Phoenix reborn from the ashes, my ordeal galvanized unyielding reflections the essence of which unveiled invaluable life lessons. Above all, I gleaned the notion that the existence of antagonizing forces shouldn't diminish the potency of my ambition. More importantly, I discovered that leadership goes beyond post ulations of titles and credentials. Revealingly, I remained a well-engaged council member, steadfastly championing worthy causes that resonated with my person.\n\nThe crowning lesson that emerged from my ordeal was gravity's rap acious generational appetite for failure, not being modestly egalitarian in its distribution. The experience in its entirety had gifted me a dimension of emotional depth that defied age, as I realised that my set back was not an expiration of my ambition or an indictment of my leadership capabilities; but a springboard into greater things. Now, with even more conviction, I am resolved in putting one foot in front of the other on the path to my goals. In summation, though my student council campaign was challenging, it proved to be a crucible of growth that refined my leadership capacity.

Tweet about GLTR
MIT-IBM Watson AI lab and **Harvard NLP**

Figure 1: Evaluation Result of Native Sample



When I was a sophomore in high school, I was determined to run for the student council president position. I had always been an active member of the council and wanted to take on a leadership role. However, my campaign did not go as planned. One of my competitors spread rumors about me, and my posters were vandalized. The negativity deeply affected me, and I eventually lost the election.\n\nAt first, I was devastated. I had put in so much effort and did not deserve the way I was treated. However, as time passed, I realized that the experience taught me valuable lessons. Firstly, I learned that there will always be people who try to tear me down, but it's important not to let their negativity affect me. Secondly, I discovered that I don't need a title to express my passion for leadership and helping others. I continued to be an active member of the student council and dedicated myself to causes that I believed in.\n\nPerhaps most importantly, I learned that failure is not the end. Losing the election didn't mean that all of my hard work was for nothing or that I wasn't fit to be a leader. It simply meant that it wasn't my time yet. I've grown from this experience and am now even more determined to achieve my goals.\n\nIn conclusion, while the obstacles I faced during my student council campaign were difficult, I'm thankful for the lessons they taught me. The experience showed me that with dedication and determination, I can overcome setbacks and become a better leader than ever before.

Figure 2: Evaluation Result of Non-Native Sample

5 Experiments

5.1 Data

The dataset we used to trained the *M_base* is H3C dataset proposed in [2]. The dataset contains tens of thousands of answer pairs from human and ChatGPT to the same questions. The dataset we used to trained the *M_1* and *M_2* are the essay dataset proposed in [3]. This dataset collected native and non-native writing samples as well as the ChatGPT answers to the same essay prompts. We listed some samples in AppendixA.

5.2 Evaluation method

To evaluate the performance of three models in indentifying AI-generated text as well as their possible bias towards non-native speaker, we choose two test dataset. One dataset is composed of TOEFL essays from non-native writer, while another dataset is composed of US 8th grade essays from native writer.The test dataset is never seen by the model when training. We used misclassified rate as our metric.

5.3 Experimental details

In all training process, we adopt same setups to control the factors.

- Batch Size = 64
- Learning Rate = 2e-5
- Epochs = 5
- Optimizer adopts AdamW as implemented in pytorch,

5.4 Results

5.4.1 Results for first hypothesis

Model	TOEFL Data	US Essay Data
M_base	56.5%	12.3%
M_1	27.2%	8.4%
M_2	54.2%	6.9%

Table 1: The misclassified rate for non-native and native samples

5.4.2 Results for second hypothesis

Model	TOEFL Data	ChatGPT-improved Data
M_base	56.5%	9.6%
M_1	27.2%	10.2%
M_2	54.2%	8.9%

Table 2: The misclassified rate for non-native and ChatGPT polished samples

6 Analysis

After training on the TOEFL essay data, the mis-classified rate1 of *M_1* decreases from 56.5% to 27.2%, which is a grand improvement. This result validates the first hypothesis. We proved that by adding non-native writing samples to training set, the bias tend to decrease. Therefore, the mismatch of training dataset can be one possible cause of the bias in GPT-Detector.
We also noticed a small decrease (from 12.3% to 8.4%)in US 8th data, which is unexpected. We suppose it indicated that the model is underfit. In another word, the potential of model is not developed fully, and adding more data to the model can improve the power of model to the next level. We also noticed that the improvement of *M_2* in US 8th data is bigger than in *M_1*, which is expected, because the *M_2* is trained on other samples of US 8th grade data, and therefore learnt more knowledge than *M_1*, which is trained on TOFEL essays.
In the second table2, there is a significant decrease of misclassification rate when identifying the essay polished by ChatGPT, which proves our second hypothesis, which is, employing advanced word choices can bypass the detector. It can be concluded that the relatively simple and repeated linguistic expressions in non-native writing samples are the cause of biased behavior made by detector.

7 Conclusion

In this project, we made two hypothesis about the cause of bias in detector and proved them with experiment. The valuable lesson we learnt from this experience is that biased training dataset can lead to biased model performance. Therefore, to make the model unbiased, we need to include training dataset from a large variety of topics. Although our model *M_1* has less bias than its predecessor *M_base*, the bias still exists. So future work can be adding more non-native data to the model. If the data is hard to acquire, some data augmentation techniques can be employed.

References

[1] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*, 2019.

[2] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*, 2023.

[3] Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. Gpt detectors are biased against non-native english writers. *arXiv preprint arXiv:2304.02819*, 2023.

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[5] Edward Tian. Gpt zero. `https://gptzero.me/`.

A Appendix (optional)

Some samples from the essay dataset are listed here:

A.1 Samples of TOEFL Essays

- I prefer to cook at home. First of all, it is cheaper and safer than the restaurant. Because I can buy the things needed for cooking, and I can make sure that they are clean and fresh. Second, making a good dinner helps me obtain a sense of achievement. On every Spring festival, I always make dinner for my whole family, they always think that the meal is delicious and we can chat freely around the table. I am really proud of it and I think it can improve the relationship between my family and me.
- The place I would like to visit most is the outer space0cthe place where most of the physical laws on the earth do not apply. It is interesting to see everything floating in the air. And the term 2̆01cin the air2̆01d should also be modified simply because there is no air anyway. It is interesting to walk the distance in a whole new style. And it is interesting to see the mother planet from a whole new angle. This big blue crystal sphere0cfrom the pictures taken by those astronauts, the sight is marvelous. And I am dying to see this big blue ball in this style.
- I imagine my life ten years in the future to be free. My life will be different in one way from now that I don2̆019t have to be tied to a job. I can travel around the world, go anywhere, see different things, meet different people, broaden my horizon and thus feel the world. I don2̆019t have to confine myself to just one place. I can go to France, South Africa, South America and Islands in the Pacific. In addition, I can be a freelancer, live on my royalties. I will have plenty of time to write novels and read all the masterpieces. This is what I imagine my life will be in ten years.

A.2 TOEFL Essays polished by ChatGPT

- I favor preparing meals at home. Primarily, it’s more cost-effective and secure than dining out, as I can purchase the necessary ingredients and ensure their cleanliness and freshness. Secondly, creating a delicious meal provides me with a sense of accomplishment. Every Spring festival, I always prepare dinner for my entire family, who consistently find the feast scrumptious and enjoy engaging in lively conversation around the table. I take great pride in this, and I believe it strengthens the bond between my family and me.
- The destination I’m most eager to explore is outer space, where many of Earth’s physical laws cease to apply. It’s fascinating to witness objects floating effortlessly, and the phrase ïn the airr̈equires adjustment since there’s no air to speak of. It’s intriguing to navigate in an entirely novel manner and observe our home planet from a fresh perspective. This vast blue globe, as captured in breathtaking photographs by astronauts, offers a truly awe-inspiring view. I’m eager to experience this magnificent sight firsthand.

- I envision my life ten years from now as one of unbounded freedom. In contrast to my current circumstances, I will no longer be tethered to a conventional job. Instead, I will embark on a journey to explore the globe, immersing myself in diverse cultures, witnessing stunning landscapes, and forging connections with unique individuals. This will inevitably expand my horizons and enable me to truly experience the world.Unshackled from a fixed location, I will traverse France, the African continent, South America, and the islands of the Pacific. As a freelancer thriving on royalties, I will have ample time to indulge my passions for writing novels and immersing myself in literary masterpieces. This is the life I foresee for myself in a decade's time.