AI-generated Text detection

Shivam Shrikant Zample.
B.Tech in Mathematics and
Computing.
IIT Ropar, Class of 2026
shivamzample@gmail.com

1 July,2024

Abstract

Ongoing development in gpt models have raised the concern about the misuse of generated texts like causing disruptions in education system,cheating writers etc.Many detection models are proposed .The major challenges they are facing include detecting paraphrased text and high false positive.This study details with possible method to tackle these issues.

1 Introduction

Large Language Models (LLMs) represent a significant breakthrough in the area of natural language processing (NLP).The ongoing advancements in the generative models keeps on reducing the gap between human and ai generated text making it more difficult to distinguish between them.

This study first examines the statistical differences between human and AI-generated text. Factors considered include entropy distribution, part-of-speech and dependency analysis, and sentiment analysis. Furthermore, we will look at how the entropy distribution can be used to detect whether the text is human or AI generated, followed by using GAN-BERT to tackle paraphrasing attacks.

2 Statistical Contrasts in Human vs AI Text

NOTE: The observations are made on the basis of this: data.

2.1 POS and dependency analysis

Figure 1 illustrates two significant observations about AI-generated texts: High noun count indicates more informativeness and objectivity which can be seen in AI generated texts.

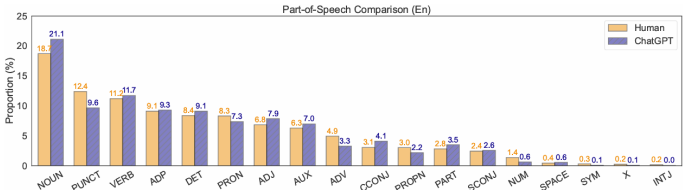


Figure 1: reference from Guo et al. [2023]

Less punctuation count indicates that GPT prefers to generate text of longer length.

2.2 Sentiment analysis

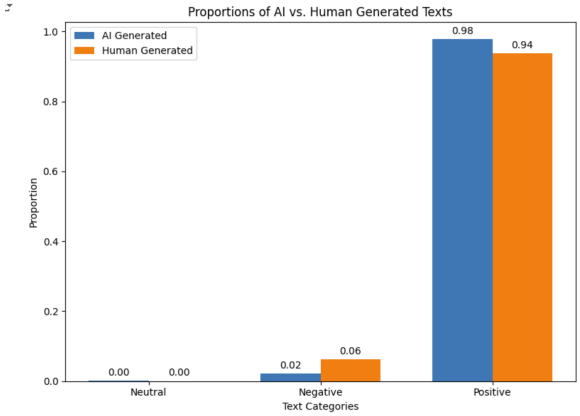


Figure 2: Observed on dataset:data

Sentiment analysis, which determines the sentiment expressed in text, plays a crucial role in identifying the text's origin. We explored semantic features: sentiment polarity (ranging from -1 to +1, indicating negativity to positivity) leveraging TextBlob for our analysis.As observed in the Figure 2, GPT generally avoids generating text of negative sentiment(as the ratio of number of negative texts to positive texts is very low as compared to that of human-generated text.)

2.3 Analyzing Text Entropy

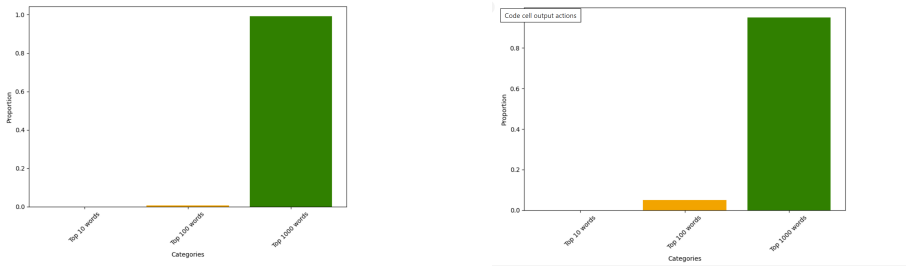
Reference from Gehrmann et al. [2019]

The main point distinguishing AI and human-generated text is their generation tecnique.Human generate text with understanding of context and audience,incorporate emotions, personal experiences into their writing. Whereas

GPT generate text based on patterns learned from vast datasets, predicting the next word or phrase based on statistical probabilities.

Consider if we are given a sequence of words $X_{1:N}$ in a text X, and we want to know whether it is human or AI-generated. Since GPT-2 117M is a standard left-to-right language model, we compute probability distribution $p_{\text{det}}(X_i \mid X_{1:i-1})$ at each position i in the text X.This can be done using 'gpt2' model and 'AutoModelWithLMHead' class of hugging face's 'transformers' library.We use two tests to assess whether text is generated .**Test1**:Look at the rank of word in $p_{\text{det}}(X_i \mid X_{1:i-1})$. **Test2**:the entropy of the predicted distribution,e.g.

$$-\sum_{w} p_{\text{det}}(X_i = w \mid X_{1:i-1}) \log p_{\text{det}}(X_i = w \mid X_{1:i-1})$$



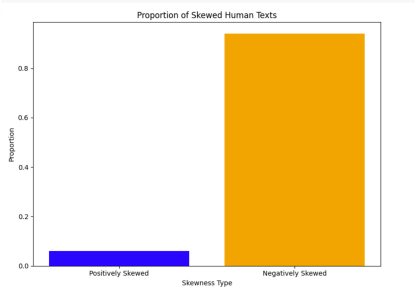
(a) human-generated text (b) AI-generated text

Figure 3: Comparison of human-generated and AI-generated texts based on samples taken from this data

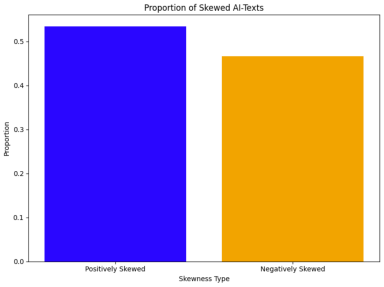
Apply Test1 to each text and according to the rank place each word of the text in one of four category namely **top-10**, **top-100**, **top-1000**,and **top-1000plus**.From the Figure 3 , we can observe that the ((top-10 + top-100) : top-1000) ratio is much higher for GPT-generated text than human one, indicating that GPT looks for high probable word that fits in the context.

After applying Test2 to each text ,we calculated the features like normalized entropy, median entropy, variance entropy, skewness entropy and kurtosis entropy for each text as seen in Figure 6 .All these features are independent of the text-length and help in capturing origin of the text.

As seen in the Figure 5 ,the entropy of human-generated text lie in higher range than AI-generated text.This is because there is no specific pattern in which human generates leading to high randomness. In contrast, GPT generates text based on patterns it has been trained on and predicting the next word based on statistical probabilities, resulting in lower randomness. The negative skewness in human-generated text (Figure 4(a))shows that bulk of the entropy values are concentrated towards the higher end of the distribution, with fewer values in the lower end, which implies most of the words in text exhibits high entropy value .Whereas many of the AI-generated text (as seen in Figure 4b) have positive

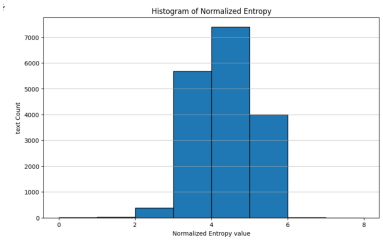


(a) human-generated text

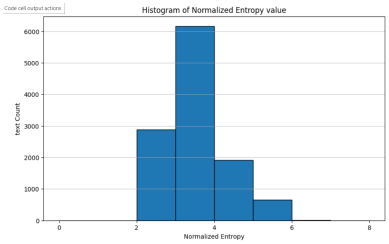


(b) AI-generated text

Figure 4: Comparison of human-generated and AI-generated texts based on samples taken from this data



(a) human-generated texts



(b) AI-generated texts

Figure 5: Comparison of human-generated and AI-generated texts based on samples taken from this data

skewness in entropy, which implies most of the words in these text have low entropy value.

normalized_entropy	median_entropy	variance_entropy	skewness_entropy	kurtosis_entropy	generated
3.781375	4.027416	2.821344	-0.342219	-0.497629	0
4.272886	4.302235	3.420157	-0.298982	-0.747111	0
4.385856	4.534761	2.636780	-0.400072	-0.160771	0
3.747545	3.899719	1.898608	-0.468827	0.391431	0
4.223971	4.260043	3.290447	-0.395210	-0.729641	0

(a) Sample of human-generated text

normalized_entropy	median_entropy	variance_entropy	skewness_entropy	kurtosis_entropy	generated
3.117894	3.151026	3.520068	0.031495	-0.953421	1
3.177591	3.354952	3.214185	-0.074241	-0.927820	1
3.165998	3.298025	3.326699	0.055291	-0.816541	1
3.141640	3.325203	3.343081	0.020164	-0.859872	1
3.091264	3.096693	3.267325	0.016501	-0.963564	1

(b) Sample of AI-generated text

Figure 6: Comparison of human-generated and AI-generated texts based on samples taken from this data, where each datapoint corresponds to a text.

3 Statistical Differences Sometimes Blur

A rephrasing and spoofing attack involves altering the text produced by an AI model to mimic or distort content. The goal is often to deceive readers or systems into believing the text is authentic.These attacks can make AI-generated text statistically resemble human-generated text.Human writing is characterized by a mix of predictability and creativity. By tuning the temperature parameter, GPT can balance these aspects to generate text that closely resembles human writing, and give entropy distribution similar to human-generated text.

To tackle this issue, we have used Generative Adversarial Networks (GANs), where my discriminator is trained on AI text that closely resembles human text.

4 Using GAN-BERT

Reference from Croce et al. [2020]

To train the model to capture mimicked text effectively, we utilize an adversarial approach with GAN-BERT.The core component of BERT is the Transformer, an attention-based mechanism that learns contextual relationships between words or sub-words (Schuster and Nakajima, 2012). BERT provides contextualized embeddings for words within a sentence and an overall sentence embedding that captures sentence-level semantics.

4.1 Method

In our study, we employ a discriminator model alongside GPT-2 from the Hugging Face Transformers library as generator ,and BERT for text analysis.
Initially, the discriminator undergoes supervised learning using labeled datasets to distinguish between human-generated and basic AI-generated text.
Subsequently, we extract the initial ten words from each datapoint containing human-generated text to grasp the discussion topic. These initial words are tokenized and used to prompt GPT-2, which then generates additional text up to a maximum length of 150 tokens. The resulting generated text is then fed into

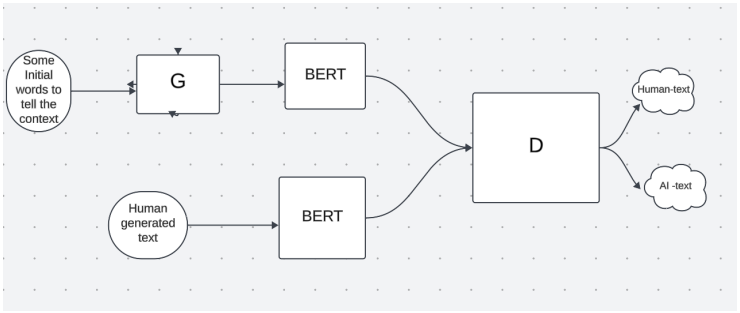


Figure 7: Architecture deployed

BERT to obtain its contextual embeddings. Simultaneously, the corresponding datapoint containing human-written text is also passed through BERT. We extract the [CLS] token embeddings from both the generated and human texts, concatenate them, and use this concatenated representation to train our discriminator model(which has already undergone supervised learning).
Why [CLS] token?
As the input tokens propagate through the transformer layers, the representation of the [CLS] token accumulates information from all tokens in the sequence. This process enables the [CLS] token to capture the context and semantics of the entire input sequence.
We define the loss functions for the generator and discriminator as follows: Let p_d and p_G denote the human data distribution and the AI-generated examples, respectively. Let us define $p_m(\hat{y} = y \mid x, y = 1)$ as the probability provided by the model m that a generic example x is associated with the AI-generated class(labelled as 1).
For the generator, the objective is to produce text that closely resembles human writing. Let $f(x)$ denote the activation on an intermediate layer of the discriminator. The feature matching loss is defined as:

$$L_{G_{\text{feature matching}}} = \|E_{x\sim p_d} f(x) - E_{x\sim p_G} f(x)\|_2^2 \tag{1}$$

This loss encourages the generator to produce examples whose intermediate representations, when input to the discriminator, closely match those of real human-written examples.

Additionally, the generator loss incorporates the error induced by AI-generated examples correctly identified by the discriminator:

$$L_{G_{\text{unsup.}}} = -E_{x\sim p_G} \log\left[1 - p_m(\hat{y} = y \mid x, y = 1)\right] \tag{2}$$

The total generator loss is : $L_{\mathcal{G}} = L_{\mathcal{G}_{\text{feature matching}}} + L_{G_{\text{unsup.}}}$
The total discriminator loss is :

$$L_D = -E_{x\sim p_d} \log\left[1 - p_m(\hat{y} = y \mid x, y = 1)\right] - E_{x\sim p_G} \log\left[p_m(\hat{y} = y \mid x, y = 1)\right]$$

This total generator loss forces gpt2 model to generate text very similar to human writing. While the total discriminator loss trains our discriminator to detect AI-generated text (upto whatever extent it mimics human-text) by capturing every minute feature.
Performing 6-8 number of epochs over entire training data, makes out discriminator ready for testing.

4.2 Result

As intially discriminator has undergone supervised learning , so the figure 8 shows its decreasing loss at every epoch.
The figure 9 shows that the generator loss keeps decreasing at every epoch, indicating that the model is improving at capturing the underlying patterns in the human-generated text data.The figure 9 also shows that the discriminator loss is also decreasing, which means it is getting better at distinguishing AI-generated text from human text with each epoch, regardless of how closely the AI tries to mimic human writing.

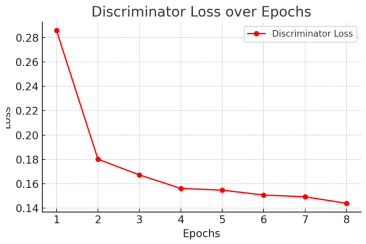


Figure 8: Supervised average training loss of discriminator

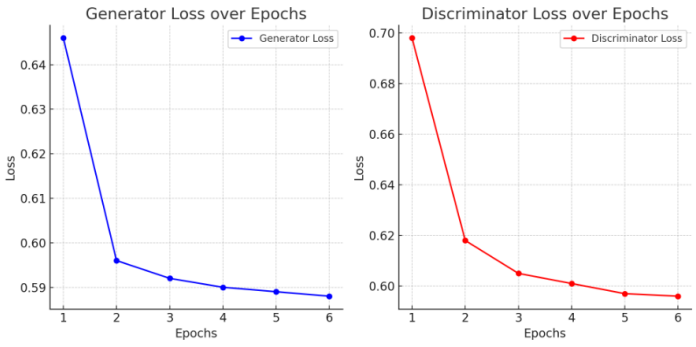


Figure 9: Unsupervised average training loss

References

Danilo Croce, Giuseppe Castellucci, and Roberto Basili. GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.191. URL `https://aclanthology.org/2020.acl-main.191`.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. Gltr: Statistical detection and visualization of generated text, 2019. URL `https://arxiv.org/abs/1906.04043`.

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, 2023. URL `https://arxiv.org/abs/2301.07597`.

4)Decoding the AI Pen: Techniques and Challenges in Detecting AI-Generated Text Sara Abdali, Richard Anarfi, CJ Barberan, Jia He

5)Decoding Text Origins: An Integrated Approach to Differentiating Human and AI-Generated Content Reddypalli Trisha, Dept Of CSE, IIIT RGUKT RK-VALLEY trishareddypalli@gmail.co

6) Working of originality ai detector, available on link

7)How to Detect AI-Generated Texts? DOI:10.1109/UEMCON59035.2023.10316132. authors:Trung Nguyen, Winona State University; Amartya Hatua, Fidelity Investments; Andrew H. Sung, University of Southern Mississippi.