	text	label
0	I rented this movie yesterday and can hardly e...	0
1	The Sentinel i was hoping would be a good film...	1
2	A top contender for worst film ever made. Joan...	0
3	Astaire and Rogers at the height of their popu...	1
4	The movie "MacArthur" begins and ends at Gen. ...	1
5	The film gives a rather condensed version of w...	0

i rented i am curious yellow from my video store because of all the controversy that surrounded it when it was first released in i also heard that at first it was seized by u s customs if it ever tried to enter this country therefore being a fan of films considered controversial i really had to see this for myself the plot is centered around a young swedish drama student named lena who wants to learn everything she can about life in particular she wants to focus her attentions to making some...

(a) IMDb dataset (Maas et al., 2011)

able	about	absolutely	acting	action	actor	actors	actress	actually	after	...
0	0	1	0	0	0	0	0	0	0	...
0	1	0	0	0	0	1	0	0	0	...
0	0	0	0	0	0	0	0	1	3	...
0	1	1	0	0	0	0	0	0	0	...
0	1	0	0	0	0	1	0	0	1	...
0	1	0	0	0	0	0	0	2	0	...
0	1	0	0	0	0	0	0	0	0	...
0	1	0	1	0	0	0	0	0	2	...
0	0	0	0	0	0	0	0	1	0	...
0	1	0	0	0	0	1	0	0	1	...

(b) Paragraph representation space

Figure 6: Figure 6(a) (left) shows examples of textual paragraphs and corresponding labels present in the IMDb dataset. It also highlights part of one random paragraph (one input) showing that in general, having a lot of sentences as input for detection is very common and practical. Figure 4(a) (right) represents the word-cloud representation of the word distribution based on which the word-level total variation is estimated. Figure 6(b) denotes the representation of the input paragraph using a Bag-of-Words-based count vectorizer for our algorithms and proving our hypothesis. It demonstrates paragraph representation space with Bag-of-Words-based count vectorizer where each row indicates one review.

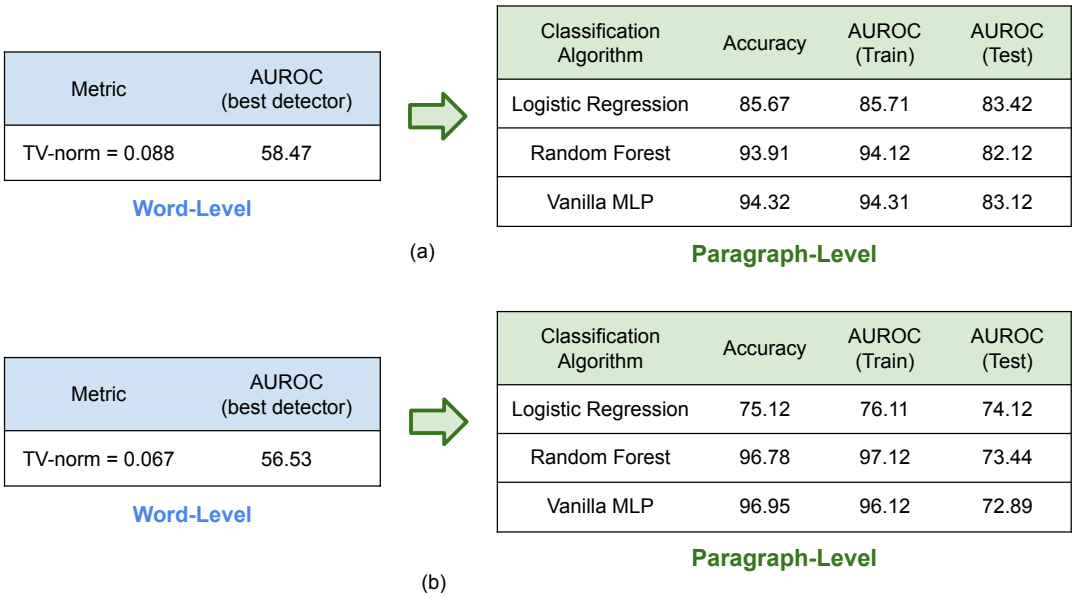

Figure 7: The table in Figure 7(a) (left) represents the total variation norm distance at a word level i.e input to the detector is the word and one needs to detect if it's a positive or negative class (human or machine in our context). It also shows the AUROC that can be achieved by the best detector based on the total variation norm as shown in (Sadasivan et al., 2023). Figure 7(b) (right) shows the accuracy and AUROC achieved by real detectors (standard machine learning algorithms) at a paragraph level, where each input to the detector is a paragraph or a group of sentences. It is evident that at a paragraph level, even a simple untuned ML detector can achieve a very high AUROC of more than 85%, which was very low at a word level. Similarly, in the tables in Figure 7(b), we observe a similar behavior as we increase the hardness of the problem by reducing the number of sentences from the passage. We note that the AUROC achieved by the real detector decreases but is still much larger than the word-level best detector's AUROC which validates our claims.

Metric	AUROC (best detector)
TV-norm = 0.102	59.73

Word-Level

Classification Algorithm	Accuracy	AUROC (Train)	AUROC (Test)
Logistic Regression	86.39	86.32	84.95
Random Forest	94.40	94.39	89.99
Vanilla MLP	96.10	97.01	89.91

Paragraph-Level

Figure 8: The table in Figure 8(a) (left) represents the total variation norm distance at a word level for the Fake News dataset (Lifferth, 2018). It shows the AUROC that can be achieved by the best detector based on the total variation norm as shown is 59.73%. Figure 8(b) (right) shows the accuracy and AUC achieved by a real detector at a paragraph level goes up to 90%, which validates our hypothesis for a general class of NLP tasks

IMDb NLP Dataset Experiments with Increased Hardness. To provide additional confirmation of the efficacy of our claim, we made the experimental setting more challenging by randomly decreasing the number of sentences in each review, making it difficult for any genuine detector or classifier to distinguish. In this scenario, we again compared the performance to the previous scenario and observed that all the methods were able to achieve a test `AUROC` greater than 0.7, which is lower than the previous case. This result supports our hypothesis that as the number of samples/sentences increases, detection accuracy improves.

We conducted a similar experiment on a Fake News classification dataset (Lifferth, 2018), and the results were consistent with our previous findings. This indicates that AI-generated text can be detected, although we need to be cautious and gather more samples as the distribution becomes closer.

We would like to emphasize that the purpose of this experimentation is to demonstrate our hypothesis regarding the feasibility of detection rather than to showcase the accuracy of classification. This is because the accuracy of classification is already well-established, with a simple pre-trained BERT-based model being capable of achieving high accuracy.

D Detailed Conclusion & Scope of Future Works

We note that it becomes harder to detect the AI-generated text when $m(s)$ is close to $h(s)$, and paraphrasing attacks can indeed reduce the detection performance as shown in our experiments. However, we assert that by collecting more samples/sentences, it will be possible to increase the attainable area under the receiver operating characteristic curve (`AUROC`) sufficiently greater than $1/2$, and hence make the detection possible. We further remark that it would be quite difficult to make LLMs exactly equal to human distributions due to the vast diversity within the human population, which may require a large number of samples from an information-theoretic perspective and provides a lower bound on the closeness distance to human distributions. Diversity could lead to realistic analysis to prove that the distributions are sufficiently separated to be detectable.

We want to emphasize that as we show detectability is always possible (unless $m = h$ in exactness), in several scenarios when m and h are very close, it might need a lot of samples to detect. However, watermark-based techniques can help address this issue by causing shifts in the distributions. The additional insights from our work could help to design better watermarks, which cannot be attacked easily with paraphrases. More specifically, it is possible to create more powerful and robust watermarks to introduce a minor change in the machine distributions, and then collecting