Artificial Intelligence Group 1

Who should be held accountable for false AI-generated content on social media or in research articles?

Image generated with Gemini

Samyuktha Nair, Arnav Nair, Advay Dinesh, Advait Baijulal

Date: 25.07.2025

Mentor: Madhuvanthi Venkatesh

1. Abstract

The increasing use of generative artificial intelligence (AI) for important write-ups and documentation raises concerns about the indistinguishability between AI-generated and human-authored content. This proliferation contributes to public distrust of online information and affects domains such as social media, journalism, and academia. One of the major concerns is the misuse of AI to produce misleading or plagiarized work, undermining academic integrity and hindering genuine skill development.

To address these issues, there has been a growing interest in the automated detection of AI-generated text using classification and watermarking techniques. In this context, this paper presents a technical study of the implementation of text classification models for identifying AI-generated content. Specifically, we evaluate the performance of four encoder-based transformer models — BERT (Bidirectional Encoder Representations from Transformers), RoBERTa (Robustly Optimized BERT Pretraining Approach), DistilBERT (Distilled BERT) and TinyBERT (Tiny Bidirectional Encoder Representations from Transformers) — trained on a common dataset. We compare their classification accuracy, efficiency and inference speed to understand the trade-offs between model size and detection reliability. This comparative study highlights the strengths and limitations of each model, enabling informed choices for efficient and accurate AI text classification.

1.1 Issues Stemming from AI proliferation

AI-generated texts can undermine academic integrity by leading students to cheat or claim that generated texts as their own. Such lack of academic integrity can also lead to fake research or fraudulent research studies. Deepfake messages can also spread false narratives about people or organizations and cause misinformation to spread. The use of AI also creates economic consequences for content creators and writers as they must work against the tide of AI which can produce texts at much faster rates than they likely can as humans. With AI's vast capabilities, it may undermine the work of original artists, or may even be used to create work that people claim as original. Both circumstances produce issues for humans by risking workplace and academic integrity. Similarly, AI-generated texts may also push false agendas or biased information that can lead to misinformation rather than providing users with objective views and facts regarding certain circumstances. With such issues at hand, the regulation of AI usage becomes crucial in maintaining trust and safety online for all users and creators.

2 Technical Background

Machine learning is a branch of Artificial Intelligence focused on creating models and algorithms that allow computers to learn data patterns and make predictions or decisions without explicit programming. In its essence, it provides machines a way to improve performance through experience, similar to the way human thinking works. Within Machine Learning, there are 3 core types of learning, including Supervised, Unsupervised, and Reinforcement learning.

- **Supervised Learning** focuses on training models using labeled data, including input and output, to predict or classify test data. The model learns to predict outcomes by comparing its output with the actual labels; by minimizing the error gap between the predictions and actual labels, the internal parameters (weights) are adjusted through constant repetition. This is commonly used for classification-based tasks, such as spam detection, and regression tasks, such as predicting house prices.

- **Unsupervised Learning** works with unlabeled data, identifying hidden patterns, grouping, and structures despite having no prior knowledge of the outcomes. Using this clustering technique, similar data points are grouped together, or dimensionality reduction occurs. The focus lies on understanding data structure to enforce predictions through feature analysis.
- **Reinforcement Learning** is based on agent interactions with a decision-making environment. It learns the optimal actions through trial and error; at the same time, feedback is given in the form of rewards and penalties. Over time, it focuses on minimizing penalties and maximizing rewards, helping in dynamic decision-making tasks such as robotics and gameplay.

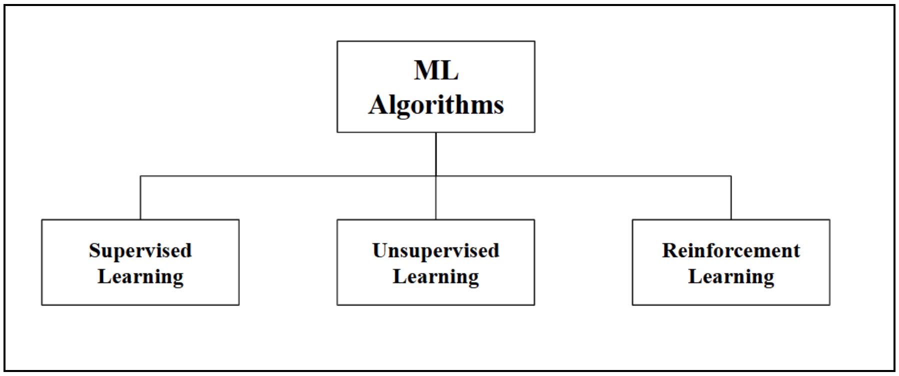


Figure 1: Machine Learning Categories

These three types of learning mechanisms provide the foundation of Machine Learning. Each technique has a unique purpose. Based on the problem that needs solving, these three types can adapt their bases to help produce results and transition into more advanced ML techniques and applications.

2.1 Transformers

Transformers are a type of deep learning neural network architecture that helps transform an input sequence into an output sequence. This has helped to revolutionize the field of Natural Language Processing (NLP). It transforms an input sequence into an output sequence based on the contextual relationship between words within a sentence, despite the position of the word. Previously, the initial model known as Recurrent Neural Networks had the issue of miscalculation due to longer sequences taking more time to calculate and limited memory capacity, which made RNNs “forget” previous data. The development of RNNs helped increase scalability massively, and the issue within RNNs was solved through the mechanism of self-attention, where the importance of each word about others was determined simultaneously. The input received by transformers is first converted to vector embeddings, often a value between 0 and 1 to reflect the similarity to other words in the sequence. This mechanism then weights the vector to determine which parts of the input are most relevant for the output. This parallel processing capability helps make Transformers much faster and effect on a larger scale all over the world, including tasks such as language generation and translation.

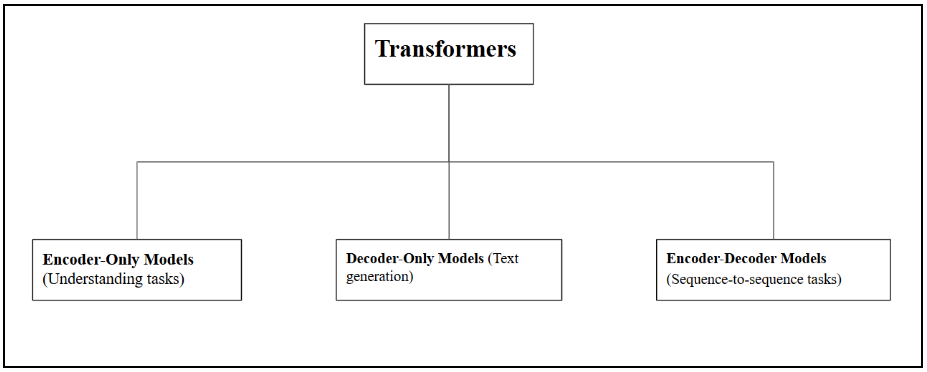

Transformers
Encoder-Only Models (Understanding tasks)
Decoder-Only Models (Text generation)
Encoder-Decoder Models (Sequence-to-sequence tasks)

Figure 2: Classification of Transformers

Among the types of Transformer Architecture existing, there are 3 main ones used in countless programs today. The 3 types include: Encoder-only, Decoder-only, and Encoder-Decoder architecture.

- **Encoder-only Architecture** focuses on reading the entire input sequence and producing a contextualized representation of each token, and with no decoder present, it isn't designed for generating new sequences. This helps with understanding-based tasks like classification and entity recognition. The process works as follows: we have an input that transitions to an encoder, which then outputs embeddings to a task-specific head. The primary example where this is used is BERT (Bidirectional Encoder Representations from Transformers)
- **Decoder-only Architecture** focuses on taking the sequence and generating the new token one at a time by reading the words individually, left to right, using self-attention. This is used for text generation tasks mainly. This works by starting with a token, then a decoder generates the 1st word, which is fed back into another word. This process continues until all tokens are used. The primary example of this is used in GPT (Generative Pre-trained Transformer)
- **Encoder-Decoder Architecture** combines both an encoder and a decoder to read and compress input into a context vector and then generate an output based on context. This is used mainly for sequence-to-sequence tasks such as question answering. This works with an input that is later given to an encoder, which generates a context vector, and this is taken to the decoder, which comes with a final output sequence. This is primarily used in T5 (Text-to-Text Transfer Transformer).

BERT is an encoder-only architecture first created in 2018 by researchers at Google AI Language. It relies on an attention mechanism that learns contextual relationships between words in a text. The goal is to generate a language representation model from only the encoder part. Its main duty is to improve performance on a variety of NLP tasks, including question answering, sentence classification, and entity recognition. BERT was a huge advancement, but now it has become even more powerful by improving models. RoBERTa (Robustly Optimized BERT approach), developed by Facebook AI, builds on BERT by utilizing more data, training it for longer periods, and applying dynamic masking, which enables higher accuracy. DistilBERT, on the other hand, is a smaller yet efficient BERT, designed for speed. It employs a technique known as knowledge distillation to compress the original BERT model. It learns based on the BERT knowledge and outputs to create a model retaining its original performance with a smaller size. TinyBERT is very similar to DistilBERT as it is the most compact version of BERT, and is usually the highest performing. In comparison to DistilBERT, the only difference is that DistilBERT is distilled with BERT-Base which is 12 layers, while TinyBERT is distilled with BERT-Base and BERT-Large (12 and 24 layers).

Table 1 BERT Model Comparison

Model	Key Innovation	Approx. Parameters	Relative Size	Relative Speed	Core Use Case
BERT-Base	Deep Bidirectionality (Masked Language Modelling + Next Sentence Prediction)	110M	1x	1x	General Purpose NLP, Fine-tuning
RoBERTa-Base	Optimized Pre-training (Dynamic Mask, No NSP, More Data)	125M	~1.1x	~1x	State-of-the-Art Performance, Research
DistilBERT	Knowledge Distillation (Triple Loss)	66M	~0.6x	~1.6x Faster	Production Systems, Real-time APIs
TinyBERT (4L)	Two-Stage, Layer-wise Distillation	14.5M	~0.13x	~9.4x Faster	Edge Devices, Mobile Applications

2.2 Type of Dataset & Label Explanations

The dataset used for differentiating between human written and AI generated text contained a list of strings with an equal representation of both listed types. Equal representation of both types of text is important as it ensures balanced learning for the model being used. Each entry in the dataset is labeled with a value of 0 or 1 meaning human written or AI generated respectively. This type of labeling is needed for binary classification tasks and for ensuring that the data is truly balanced to prevent any introduction of bias during training. By working equal proportions of both types it minimizes skewed predictions but also enables a more accurate evaluation of its true performance across both classes. The full dataset contains 20,283 entries, providing a substantial volume of examples for both training and evaluation. The methodological choice supports model validation through metrics like accuracy, precision, recall, and F1 score which can be meaningfully interpreted only when both classes are adequately represented. This is relevant in the real world as distinguishing between human written and AI generated texts are becoming increasingly important. The datasets were taken from Kaggle [4, 5].

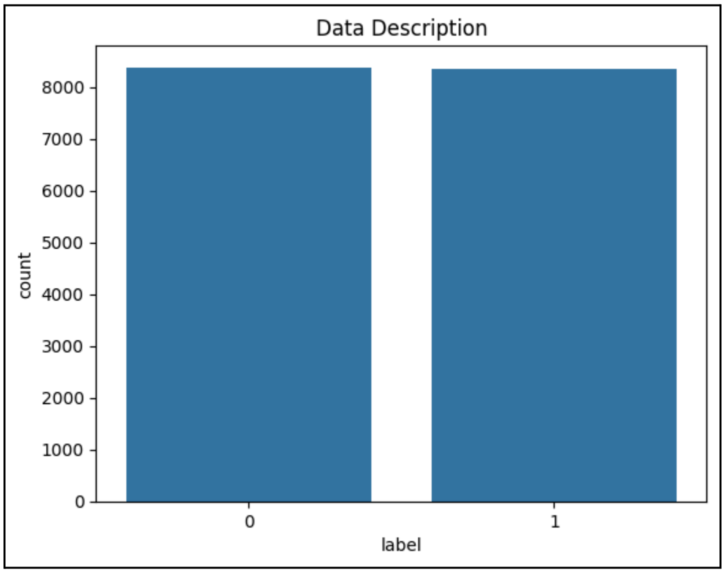


Figure 3: Dataset label distribution against total count.

2.3 Filtering & Preprocessing Methods

Filtering and preprocessing are ways to improve the consistency of the output and are necessary before feeding the model any data. The steps in this process included converting all the text to lowercase to remove any case related inconsistency which could otherwise lead to redundant token expressions (“The” vs “the”), and removing punctuation to focus on content rather than formatting. Then non alphabetic words were filtered out to reduce disorder from symbols that may confuse the model. Additionally, stop words were applied and eliminated commonly used but low information words, allowing the model to concentrate on the most meaningful features of the text. This step helps emphasize more informative features of the text that are more likely to reveal structural or patterns unique to either type of text. The filtering and preprocessing functions aimed to standardize the input, reduce irrelevant variation, preserve linguistic cues relevant to the classification, and therefore improve the model’s ability to learn generalizable patterns that distinguish human written and AI generated text. To achieve this, the Natural Language Toolkit (nltk) python package was used.

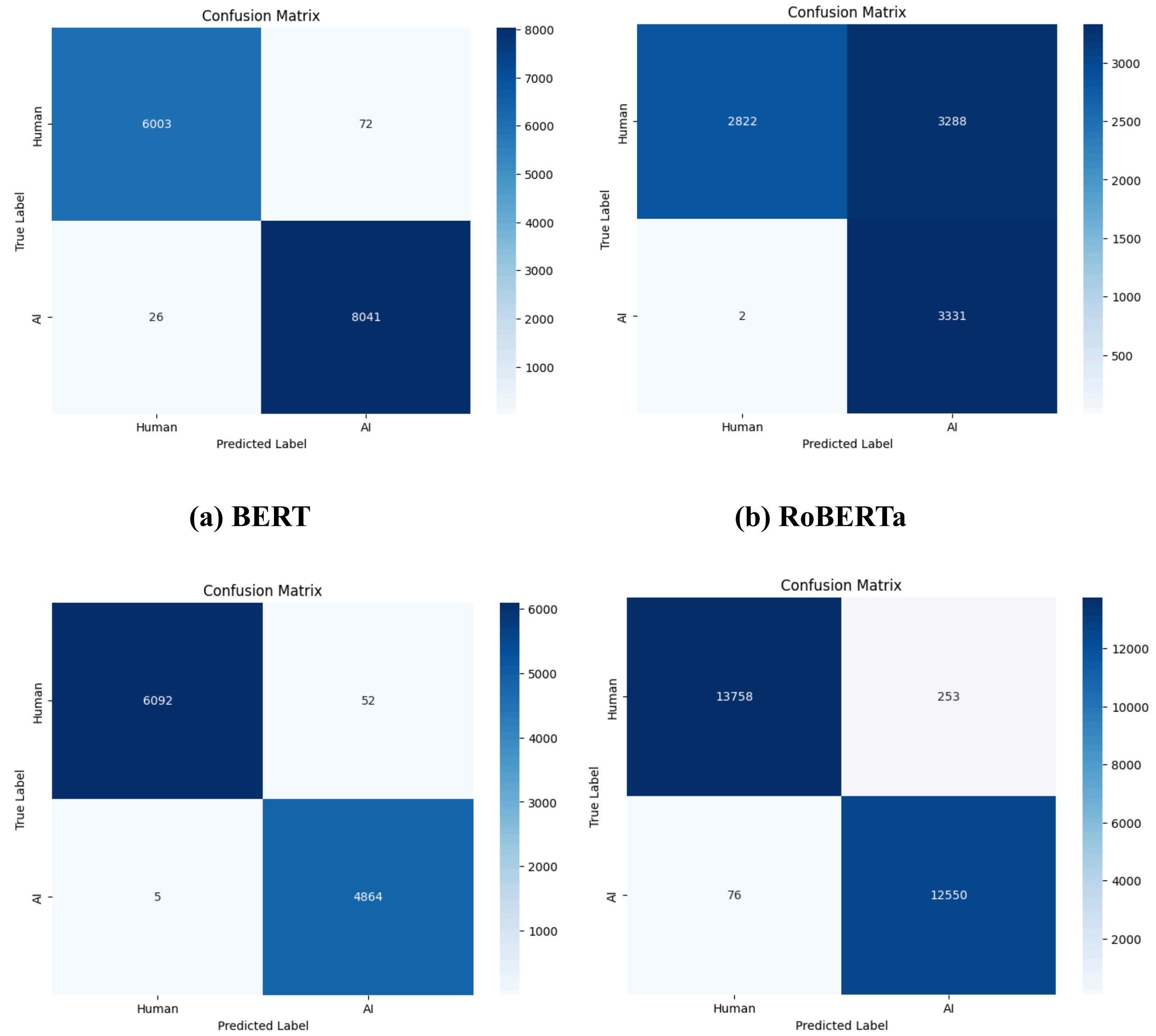
3 Results

From our research and testing using a variety of sample texts from a big paragraph to only 2-3 sentences, it is visible that all the BERT models had an easier time classifying the bigger pieces of text in comparison to the smaller text. The 3rd sample is a case where the shorter paragraph caused the model to predict humans despite it being made from AI. This struggle is primarily caused by BERT's bidirectional processing capabilities. It excels in longer texts as it can find the context more easily due to a higher word count; however, with a lower word count, it struggles to find small nuances or signals to properly predict who created the text. However, in most cases, AI texts were harder to identify than human texts. In the case of human texts, the detection was much better.

Table 2 Results

Sample Text	**Actual Classification**	**Model Prediction**
Global warming is a serious threat today. We need to take serious action to promote sustainability and address climate change. In the long term, this could pose a serious threat to our environment and many species.	Human	Human
Climate change has become an indigent topic in our day-to-day life. From bush fires in New South Wales and Victoria to the wildlife forest fires in California, it has become an inevitable subject that needs more attention. One of the vital causes of climate change is the disappearance or extinction of important species. The species could range from insects to wild mammals. Hence, it is crucial to trace the location species that play a vital role in the food chain and ecosystem to further study and conserve. We introduce a novel method to extract the geographical coordinates of localities and the distribution of species from georeferenced maps. This is done through a series of processing steps. We propose a software toolbox that extracts maps from the textbooks,occurrence points (distribution points of the species) from these maps and finally, georeferencing and postprocessing the maps to extract the geographical coordinates of the occurrence points. The data relating to the distribution of species, their habitats would pave the way for the study of functional traits or habitats of species with their abiotic properties of the environment (Zeuss, 2020).	Human	Human
The CPU (Central Processing Unit), often referred to as the "brain" of the computer, is a critical hardware component that performs most of the processing inside a computer. It executes instructions from programs by performing basic arithmetic, logic, control, and input/output (I/O) operations.	AI	Human
Of course. Let's place the idea of nature into a more specific context, using the current time and location. Here in Offenbach, Germany, on a warm afternoon in early July, nature feels both immediate and serene. The sun is likely casting a warm, golden light over the city, coaxing the deepest greens from the leaves of the trees lining the Main River. This time of year, parks like the Büsing-Park or Dreieichpark are in their summer glory, with flowerbeds bursting in vibrant color and the air filled with the gentle hum of bees. The experience of nature here isn't one of vast, untamed wilderness, but of its intimate integration with city life. It's found in the shade of a mature linden tree, the gentle flow of the river, and the chorus of birdsong that persists over the urban rhythm—a welcome, living counterpoint to the stone and asphalt of the city.	AI	AI

3.2 Confusion Matrix



(a) BERT

(b) RoBERTa

(c) DistilBERT

(d) TinyBERT

Figure 4: Confusion Matrix of BERT, RoBERTa, DistilBERT, TinyBERT

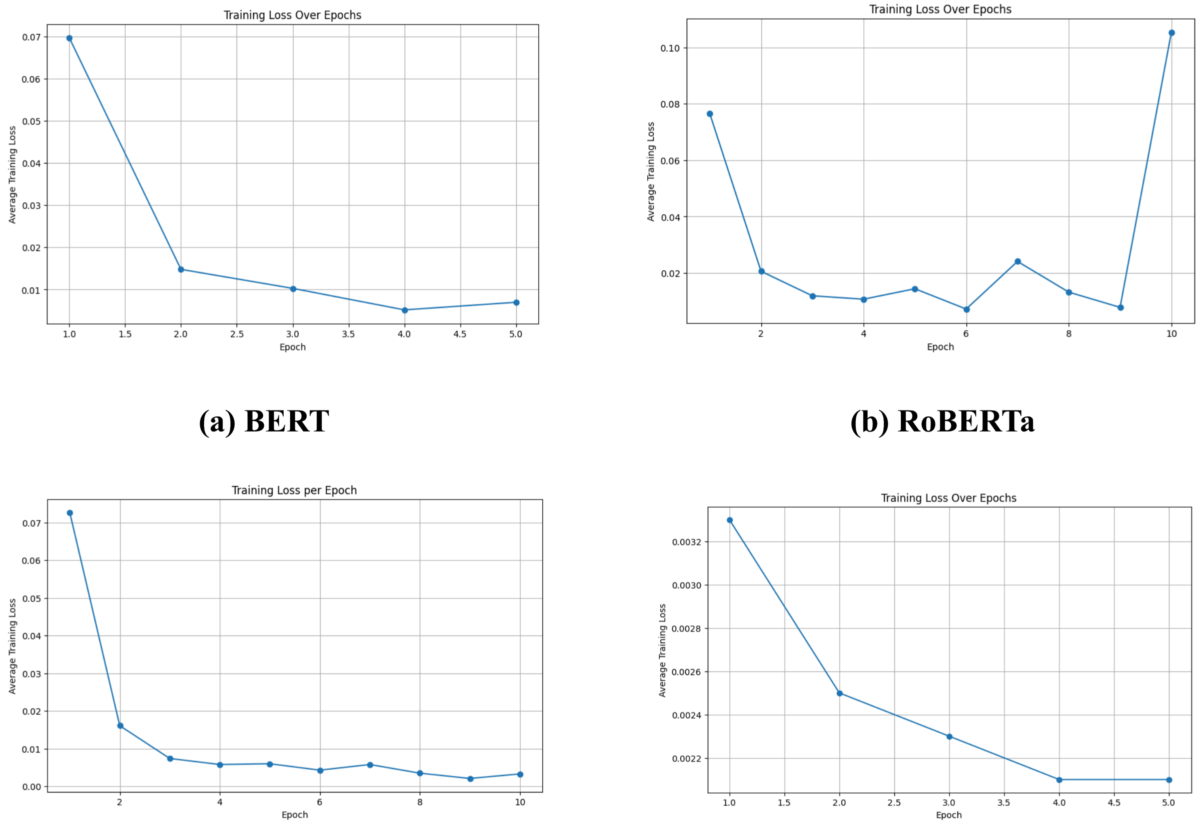
The confusion matrices in Figure 4 illustrate the classification performance of four transformer-based models: BERT, RoBERTa, DistilBERT, and TinyBERT. Each matrix shows the number of correctly and incorrectly classified samples across the two classes: Human and AI.

- **BERT** achieved high accuracy with 6,003 true human predictions and 8,041 true AI predictions, while making only 72 false positives (human misclassified as AI) and 26 false negatives (AI misclassified as human). This indicates strong overall performance with excellent precision and recall.
- **RoBERTa**, while achieving very high recall for the AI class (only 2 false negatives), showed a significant drop in precision due to 3,288 false positives. It frequently misclassified human inputs as AI (only 2,822 correct human predictions), suggesting a tendency to over-predict the AI class.
- **DistilBERT** performed robustly, with 6,092 true positives for the human class and 4,864 for AI, and very few errors: just 52 false positives and 5 false negatives. Despite being a smaller and faster model, its performance was on par with full-size BERT, indicating good balance between efficiency and accuracy.
- **TinyBERT** displayed impressive scale, processing larger data volumes (likely due to repeated input chunks caused by its limited capacity). It predicted 13,758 humans correctly and 12,550 AIs correctly, with 253 false positives and 76 false negatives. Despite its compact size, it maintained strong overall accuracy, though slightly behind DistilBERT and BERT in precision.

Overall, BERT and DistilBERT exhibited the best balance between precision and recall. TinyBERT also performed well, particularly considering the constraints on input length (128 tokens, processed in two passes). In contrast, RoBERTa, while showing high recall for AI, suffered from a high false positive rate, making it less suitable for tasks where misclassifying humans as AI is costly.

3.3 Training Losses

The Adam optimiser was used to measure the learning rate and accuracy, and the training loss was measured using the cross-entropy loss function. Cross-entropy loss is a widely used loss function for classification problems. It measures the difference between the predicted and true probability distributions of the labels for each BERT model. Cross-entropy loss penalizes the model more when it is confident about a wrong prediction and less when it is confident about a correct prediction. The Adam optimiser uses this error signal, in the form of gradients, to efficiently update the model's parameters and reduce error, thereby improving performance. A higher loss means the model's predictions are poor for that batch of data.



(a) BERT

(b) RoBERTa

(c) DistilBERT

(d) TinyBERT

Figure 5: Training Loss of BERT, RoBERTa, DistilBERT, TinyBERT

Given the graphs of the 4 BERT models used, it can be seen that as the epochs increased, naturally, the average training loss decreased as the model went through each epoch. Due to the backpropagation, more data was input to the model so it could learn how to change the weights to minimize the loss function as much as possible. To input the data into the model multiple times and constantly update the parameters, AI uses something called an epoch, which is a complete pass through the entire dataset. From here, the calculated error is updated as the epochs increase, but sometimes having too many passes may confuse the model.

In terms of irregularity, it can be seen that RoBERTa has a huge increase in training loss near 9 epochs, which is not typical. Meanwhile, DistilBERT has fluctuations starting from the 4th epoch onward, with it either increasing or decreasing. In the case of the huge spike in RoBERTa this may have occurred due to batch size issues, where there is an imbalance in the training process due to not having an integer batch size. Furthermore, the most likely outcome may be that the model is overfitting the function, and when it encountered data not following the previous trends, the training loss spiked. Similarly, DistilBERT had a small sawtooth Pattern training loss pattern from the 4th epoch onward, which could be due to a high learning rate that overshoots the minimum by just a little, causing it to increase. Then, as the model adjusts to its previous minimum, it decreases and then continues in a zigzag pattern. In general, TinyBERT, BERT, and DistilBERT had overall low training loss. RoBERTa had a comparatively high training loss in comparison. For BERT, the maximum training loss was 0.0697, while the minimum was 0.0052. RoBERTa had a maximum of 0.1054 and a minimum of 0.0072. DistilBERT had a maximum of 0.0727 and a minimum of 0.0021. TinyBERT had a high of 0.0033 and a low of 0.0021. Overall, apparently the lowest value was from TinyBERT and DistilBERT, both at 0.0021, while the highest value was from RoBERT, at 0.0766.

3.4 Evaluation

Table 3 Performance Comparison of Transformer Models on Classification Task

Model	AverageTraining Time for each epoch	Total number of epochs	Resources used	Accuracy	Precision	F1-Score	Recall Score
BERT	01:22:17	10	Tesla 4 GPU	0.9931	0.9911	0.9968	0.9939
RoBERTa	00:39:04	10	Tesla 4 GPU	0.6516	0.5032	0.9994	0.6694
DistilBERT	00:39:07	10	Tesla 4 GPU	0.9948	0.9894	0.9990	0.9942
TinyBERT	00:03:10	5	Tesla 4 GPU	0.9876	0.9802	0.9940	0.9871

3.4.1 Metrics

- **Accuracy** - In Machine Learning, accuracy is a metric that signifies the correctness of a model. This means the percentage of the model's predictions that are true positives or true negatives.

$$Accuracy \ = \frac{(True\ Positives + True\ Negatives)}{(True\ Positives + True\ Negatives + False\ Positives + False\ Negatives)}$$

- **Precision** is the measurement of how often the model's positive predictions were correct. In this case, precision is how often the model was correct when it predicted that AI created a text.

$$Precision \ = \frac{True\ Positives}{(True\ Positives +\ False\ Positives)}$$

- **Recall** is a measurement of the model's ability to detect all relevant instances in a dataset. In this case, it was the number of AI texts the model detected out of the total number of

AI texts. Recall is found by dividing the true positives by the (true positives + false negatives).

$$Recall\ = \frac{True\ Positives}{(True\ Positives +\ False\ Negatives)}$$

- The **F1 score** is the harmonic mean of precision and recall. This metric is especially useful in situations with uneven class distribution, where one class has many more instances than the other. A high F1 score indicates that the model has good precision and recall.

$$F1\ Score\ = \frac{2\ (Precision\ x\ Recall)}{(Precision +\ Recall)}$$

3.4.2 Evaluation Explanation

All models were trained on Google Colab Pro using a Tesla T4 GPU, with the tqdm package to monitor training progress. Among the four models tested, DistilBert achieved the strongest overall performance: an accuracy of 0.9948, precision of 0.9894, recall of 0.9942, and an F1-Score of 0.9990, while maintaining a relatively efficient average training time of 00:39:07 over 10 epochs. BERT also performed strongly with an accuracy of 0.9931, precision of 0.9911, recall of 0.9939, and an F1-Score of 0.9968; However, it had the longest average training time at 01:22:17 over 10 epochs. TinyBERT, though having marginally lower values, still performed strongly with an accuracy of 0.9876, precision of 0.9802, recall of 0.9871, F1-Score of 0.9940, and while training with an average training time of 00:03:10 over 5 epochs. This speed is due to TinyBERT's small architecture, which requires fewer parameters and resources to train. TinyBERT was limited to 5 epochs to avoid overfitting, a problem observed with RoBERTa. Despite having a relatively fast training time of 00:39:04 per epoch and an F1-Score of 0.9994 delivered disappointing results: accuracy of 0.6516, precision of 0.5032, and a recall of 0.6694.

The unusually high F1-score, combined with low precision, suggests prediction skew or misclassification patterns, highlighting RoBERTa's difficulty with recall and overall reliability on this task.

4 Conclusion

Over the past two decades, the advancement of generative AI has led to indistinguishability between AI and Human written text, raising concerns all across the internet. Through this study, 4 transformer-based models - BERT, RoBERTa, distilBERT, and TinyBERT were assessed on their ability to determine whether text was human or AI written. Our findings suggest that TinyBERT and DistillBERT balanced high accuracy with reduced training times, proving them ideal for real-time applications. However, BERT and DistillBERT produced the highest overall classification performance, whereas RoBERTa was the worst performing. Our study suggests that AI-text detection is possible with appropriate models, however problems still exist: especially with smaller text input. As time passes, content authenticity will continue to improve as AI-detection technology improves, fostering accountability and content integrity on the internet.

References

Alla, S. (2021, April 9). *Attention Mechanisms With Keras | Paperspace Blog*. Paperspace by DigitalOcean Blog. https://blog.paperspace.com/seq-to-seq-attention-mechanism-keras/

Allen, L. (2023, July 5). *Metacognition and self-regulation. It sounds horribly complicated, doesn't it? It's the sort phrase, which when uttered during staff meetings, causes educators to yawn, panic or roll their eyes.* Linkedin.com. https://www.linkedin.com/pulse/metacognition-self-regulation-perplexing-phrase-classroom-allen/

Gou, J., Yu, B., Maybank, S., & Tao, D. (2021). Knowledge Distillation: A Survey. *Int. J. Comput. Vis.* https://doi.org/10.1007/s11263-021-01453-z

nirmalgaud. (2024, January 27). *Human vs AI Text*. Kaggle.com; Kaggle. https://www.kaggle.com/code/nirmalgaud/human-vs-ai-text

Theocharopoulos, P. C., Anagnostou, P., Tsoukala, A., Georgakopoulos, S. V., Tasoulis, S. K., & Plagianakos, V. P. (2023). *Detection of Fake Generated Scientific Abstracts*. https://doi.org/10.1109/bigdataservice58306.2023.00011