AI Text Classification Using Ensembled Transformer Models

Brian Cong
Department of Computer Science, Eastern Michigan University
COSC480: Special Topics — Deep Learning
Dr. Ourania Spantidi
December 13, 2023

Abstract

With the widespread proliferation and ease of access of large language models like ChatGPT, the need for discriminating between AI generated and human models is now a more critical task than ever. This paper outlines a potential approach involving enhancing current state of the art architectures with human generated classification features.

Introduction

It would not be an exaggeration to say that the field of machine learning is currently advancing at an unprecedented rate. A scant six years ago, the first paper on neural transformers, *Attention is All You Need* (Vaswani et al., 2017) was released at NeurIPS 2017; in only 6 years we have gone from a single paper proposing a theoretical framework to major firms worth billions of dollars, multiple competing transformer architectures, and — as generative transformer models have proliferated rapidly, becoming accessible to the average person — AI, Machine Learning, and ChatGPT becoming household names.

This is in spite of the fact that, less than a decade ago, the concept of 'AI' was essentially unknown to the average person outside of movies like *Terminator* or *The Matrix*. But the ease of use of OpenAI's 'low-key research preview' of ChatGPT (Karen Hao, 2023) combined with the accuracy and utility of its responses were enough to make it into not only a household utility, but have even been effective enough to begin to cause major shifts to the economy. (Chui et al., 2023)

The speed at which AI text generation models have progressed has led to a capability gap: namely, that there is no widely accepted method of determining whether or not text has been generated by a large language model (LLM) or not. For some this is not a concern; to an editor, as long as a journal article generates clicks and viewership, it is largely unimportant who — or what — wrote the text. But for the readers it may be important to know that the source of an article is a human, not a machine attempting to influence public opinion. For educators it is crucial that the work done can be, in fact, verified as work done by the student, and not by an AI. And while it may be that large language models like ChatGPT eventually become so good that it is entirely impossible to determine the authenticity and origins of *any* text, that does not mean that authenticity is not still an important factor in the studies, journalism, books, and other writing that we consume.

In this paper I will compare two different state of the art models in discriminating between AI generated and human written academic texts, as well as review an aggregate ensemble model that uses both of their guesses to produce an even more accurate result.

Background

Due to the novelty of large language models and their rate of proliferation, the amount of research done on their effect upon student academic integrity, and their effect upon media authenticity is as of yet not well researched. As such, the rate of error in discriminating between AI-generated and human written text is difficult to gauge in real world terms; this is a rapidly developing situation. While a .95 AUC return on a model might be considered 'good' for some fields, in other cases a .05 rate of error might be considered unacceptable, or even high enough to render the model unusable in real-world settings. What is more likely, however, is that AI detection of AI generated texts will not be used as a final validation at any point; much like how self-driving vehicles are still required to be crewed by a person at the wheel today, the final decision maker will always be a human. As such, the accuracy rate need only be accurate enough for a user to prompt further investigation.

Currently, while there is no singularly accepted method of detecting AI-generated text, there are several differently widely used types of models. Theoretically, any model capable of performing a binary classification task on a textual corpus may be used to classify text into any two categories — this includes methods such as naive Bayes, or support vector machines, that attempt to classify texts based upon the words within them. (Mirończuk & Protasiewicz, 2018) One particularly interesting method, however, relies upon the ability of transformers themselves to generate text embeddings. As described in the original paper by Vaswani et al., transformers are capable of embedding relational information into quantized vectors. This means that, unlike the other methods currently used for textual classification and analysis, they are able to map the relationships of the words within the text to each other, rather than simply counting the number of occurrences of the word, meaning they can capture nuance that other models are incapable of doing so.

However, transformer models do not capture every possible facet of a text; for example the variant transformer architecture proposed in *Burst and Memory-aware Transformer: Capturing Temporal Heterogeneity* (Lee et al., 2023) modifies the transformer architecture so that it can capture temporal data. As such, ensembling models that capture other features that transformers do not can allow us to better classify texts in ways that a purely transformer based architecture cannot.

Model Architectures

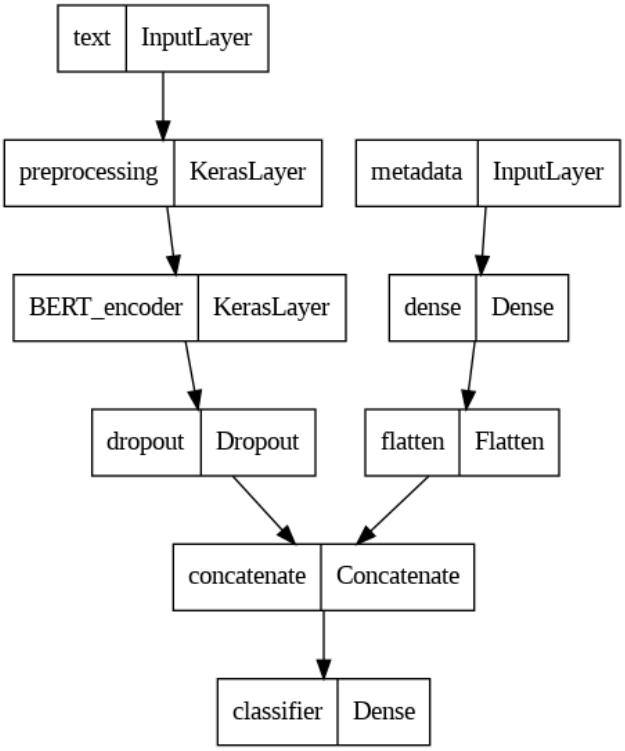
The models ensembled into the final product were drawn from two primary sources: a transformer model, *Bidirectional Encoder Representations from Transformers* (Devlin, Chang, Lee, & Toutanova, 2019), often referred to as BERT, as well as a rule-based classifier drawn from the paper *Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools* (Desaire et al., 2023). Specifically this rules-based classifier was chosen because of its high performance despite being less computationally complex, as well as the features it captures — several of the features captured

by the model would not be included in many other approaches; specifically, uses of punctuation, as well as sentence length and burstiness, are features not captured by other embedding methods.

Rules-based Classifier: While the rules based-classifier is relatively well understood, the performance outlined in the paper by Desaire et al. is far higher than expected; many other comparable algorithms, such as the methods outlined in *Distinguishing Human Generated Text From ChatGPT Generated Text Using Machine Learning* (Islam et al., n.d.), achieve a far lower accuracy rate of .75 - .80 AUC, whereas the method outlined by Desaire et al. achieves an accuracy of .91 - .93 AUC at the paragraph level by aggregating over 20 different metrics. Most notably, while other methods capture primarily word embeddings, Desaire et al. go out of their way to include punctuation and temporal data into their model features.

Transformer Model: The bidirectional transformer model, BERT, is an improvement over the baseline transformer model and does so by generating embeddings from both directions of a text. (Devlin, Chang, Lee, & Toutanova, 2019) While the potential of transformers is not yet fully understood, the transformer architecture, at a base level outlined in the original paper by Vaswani et al., does not have the ability to capture punctuation or temporal information, both of which are key metrics in classifying human-generated text from AI generated text. (Ek, Bernardy, & Chatzikyriakidis, n.d., Lee et al., 2023)

Fig 1 Final ensembled model with bidirectional transformer and rules based classifier



Full model with feature vectors available in Fig. 4

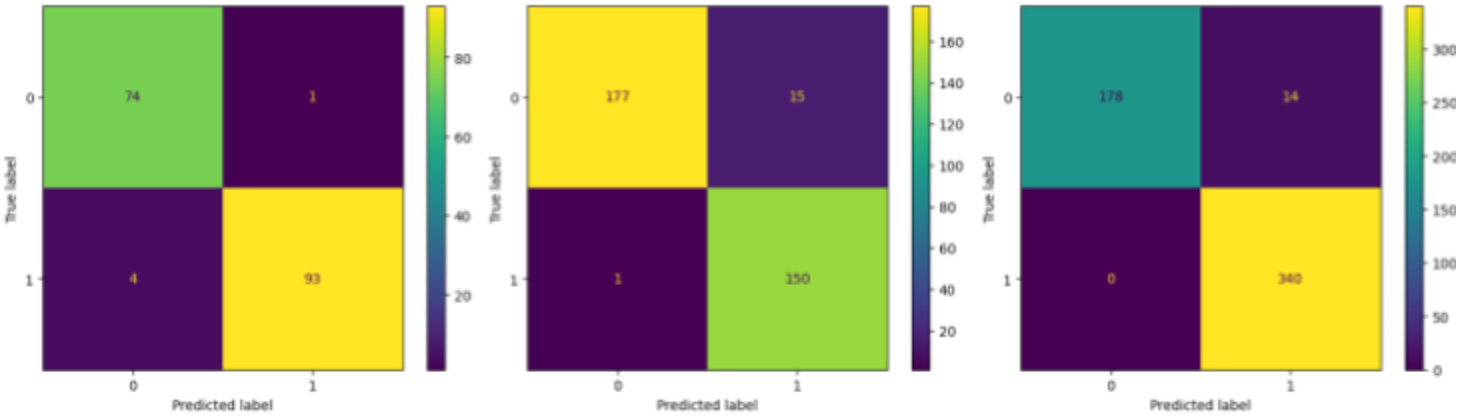
Data Augmentation

The original dataset provided in the study by Desaire et al. had a relatively small number of examples. To increase the amount of training data as well as test and validation data for higher fidelity in results, OpenAI's ChatGPT API was queried to build another dataset, using the same queries that generated the original dataset. As LLMs are non-deterministic, the texts generated by GPT-3.5 varied from those given for the original dataset, and so were appended to the original dataset provided by Desaire et al. to augment the existing data.

Results

The original rules-based classifier was the least accurate, with the AUC being .93. The purely BERT model, using a relatively small set of pretrained weights that were then tuned, resulted in an AUC of .95, while the ensembled model provided a final AUC of .97, a significant increase over the baseline BERT model.

Fig. 2 Confusion Matricies of Models

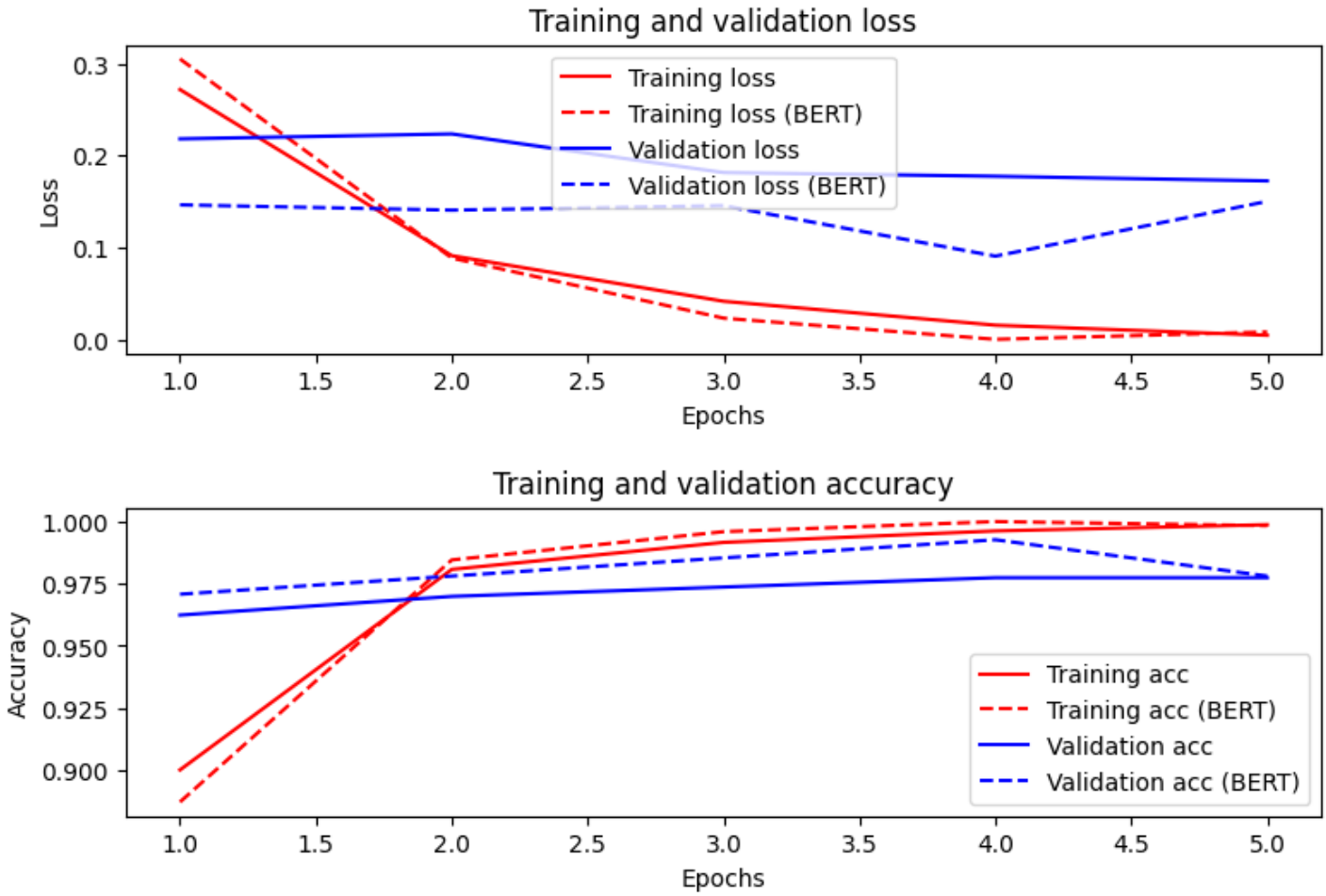


Rules Based Model *BERT* *Ensembled Model*

The labels used were 0 for human generated, and 1 for AI generated. The matrices indicate that, even as the datasets are augmented in size, the number of examples that receive false positives are roughly the same, while the number of samples that receive false negatives actually decrease, which reflects the final AUC and accuracy scores increasing as well.

As shown in Figure 3, the training and validation loss do not measurably increase against BERT; this is likely due, firstly, the training accuracy already being very high, and secondly, the dataset may be too small for such an advanced model like BERT — even with a smaller set of pretrained weights, most likely what is happening is that it is overfitting to the available training data to an extent; as such, further data augmentation will likely yield more information and further results as to whether that is happening.

Fig. 3 Training and validation loss per epoch comparison, BERT versus Ensembled BERT



Discussion

While this particular paper should not be regarded as a comprehensive study upon the shortcomings of transformer models or serve as a state-of-the-art basis for AI text detection, the lack of other AI text detection research means that, as of right now, this seems to be the most effective result with regards to AI detection research. Other detection systems, such as GPTZero, have no publicly available research into their methodologies; what available information is available suggests that they use some similar metrics as to what is explored here — their help article suggests that they likely use a burstiness and perplexity metric to output their final score, which is similar to some of the metrics used by Desaire et al. in their XGBoost based model. (GPTZero, n.d.)

Further research upon this subject should include comparisons to different ensembled features — a ‘bag-of-characters’ approach, as well as the use of the memory-aware transformer model could inherently capture more information than the model produced here can. Additionally, it is crucial that a larger set of examples be produced, and not only from GPT-3.5, but also from GPT-4.0. It may be that models discriminating between 3.5 and 4.0 need to even be

separately trained, as the different model weights may result in entirely different types of features that can be used to differentiate between the two and humans.

Additionally, different prompts should be used; a singular prompt was used here, as research limitations prevented a thorough exploration of different outputs from different prompts. Prompt obfuscation may be a problem that real-world applications need to contend with; while it is possible to preclude humans attempting to write like AI by selecting for domains, it may be that in certain cases people using LLMs attempt to obfuscate their work using prompt engineering.

Conclusion

While the arms race between generative AI and classification models has been continuous since the introduction of ChatGPT to the public, and will continue for the foreseeable future, it is possible today to discriminate between AI generated text from GPT-3.5 and human written academic texts at a reasonable rate, and models doing so can be used to serve as preliminary detection systems for academic dishonesty or media authenticity. Generative AI will typically have different priorities and will see texts differently than humans. By understanding the gaps in a model's viewpoint and reasoning, we can then develop models that seek to exploit them in order to detect AI generated content.

Additional Figures

Fig. 4 Full ensembled models with inputs and outputs

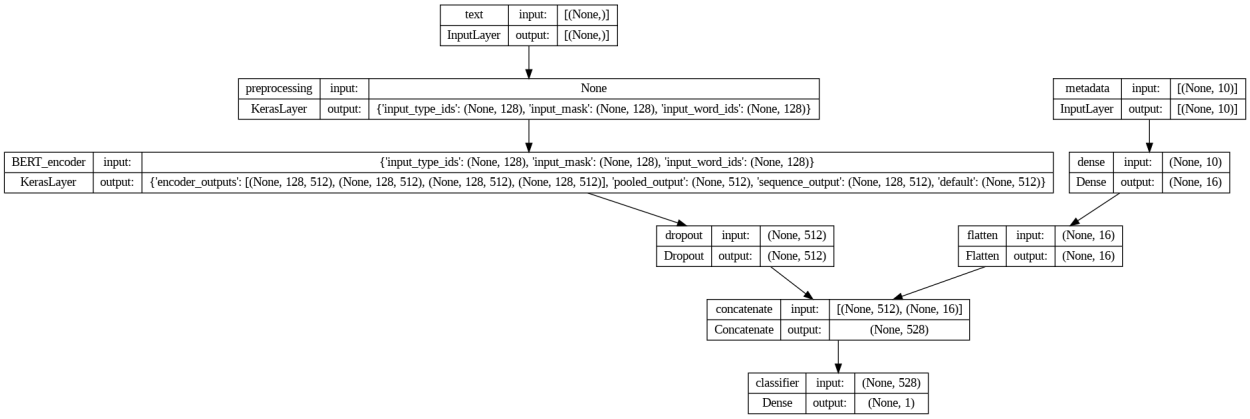


Table 1

Features included from XGBoost Model developed by Desaire et al.

Feature Number	Short Description	Greater In
2	Sentences per paragraph	human
3	“)” present	human
4	“-” present	human
5	“;” or “:” present	human
6	“?” present	human
7	“‘” present	ChatGPT
8	standard deviation in sentence length	human
9	length difference for consecutive sentences	human
10	sentence with <11 words	human
19	sentence with >34 words	human

Bibliography

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023, August 2). *Attention is all you need*. arXiv.org. https://arxiv.org/abs/1706.03762

Introducing chatgpt. Introducing ChatGPT. (n.d.). https://openai.com/blog/chatgpt

Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zemmel, R. (2023, June 14). *The economic potential of Generative AI: The Next Productivity Frontier*. McKinsey & Company. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#introduction

Karen Hao, C. W. (2023, November 30). *Inside the chaos at OpenAI*. The Atlantic. https://www.theatlantic.com/technology/archive/2023/11/sam-altman-open-ai-chatgpt-chaos/676050/

Mirończuk, M. M., & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, *106*, 36–54. https://doi.org/10.1016/j.eswa.2018.03.058

Lee, B., Lee, J.-H., Lee, S., & Kim, C. H. (2023). Burst and memory-aware transformer: Capturing temporal heterogeneity. *Frontiers in Computational Neuroscience*, *17*. https://doi.org/10.3389/fncom.2023.1292842

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *In Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)* (pp. 4171-4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Desaire, H., Chua, A. E., Isom, M., Jarosova, R., & Hua, D. (2023). Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools. *Cell Reports Physical Science, 4*(6), 101426. https://doi.org/10.1016/j.xcrp.2023.101426

Islam, N., Sutradhar, D., Noor, H., Raya, J. T., Maisha, M. T., & Farid, D. M. (n.d.). Distinguishing Human Generated Text From ChatGPT Generated Text Using Machine Learning. Department of CSE, United International University (UIU); Department of CSE, University of Asia Pacific (UAP), Bangladesh. Email: [nislam201057@bscse.uiu.ac.bd,

dsutradhar201046@bscse.uiu.ac.bd, hnoor222007@mscse.uiu.ac.bd, 20101002@uap-bd.edu, 20101001@uap-bd.edu, dewanfarid@cse.uiu.ac.bd].

Ek, A., Bernardy, J.-P., & Chatzikyriakidis, S. (n.d.). How does Punctuation Affect Neural Models in Natural Language Inference. Centre for Linguistic Theory and Studies in Probability, Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg. Email: [adam.ek, jean-philippe.bernardy, stergios.chatzikyriakidis@gu.se].

GPTZero. (n.d.). How do I interpret burstiness or perplexity? Retrieved from https://support.gptzero.me/hc/en-us/articles/15130070230551-How-do-I-interpret-burstiness-or-perplexity-