Multigenerator and Multidomain Machine-Generated Text Detection

Abstract

Introduction

Several issues are brought up by the widespread use of AI language models, the most important of which being the possible compromise of intellectual and academic integrity. It's becoming more and more important to be able to tell the difference between material written by AI and human writers as AI-generated content and human-authored works. It is crucial to develop techniques for differentiating content produced by AI from that authored by human writers since the spread of AI-generated information may unintentionally compromise the legitimacy and authority of intellectual works. It has been suggested that machine learning may be used to identify text producers. This study assesses the efficacy and precision of several machine learning techniques for distinguishing ChatGPT-generated text from human-generated material. The machine learning algorithms are given the classification job, and a thorough performance comparison is carried out. This research addresses the immediate problem of differentiating text generated by ChatGPT from human-written content through the development and evaluation of our machine learning system. Additionally, it broadens our understanding of potential future locations and mitigation strategies for AI-produced material. In summary, in an era of fast evolving natural language generation models, our research has important implications for upholding the moral use of AI technology and the reliability of digital communication.[10].

The simplicity of using massive language models (LLMs) is making machine-generated material more commonplace in a range of contexts, including education, social media, news, forums, and even academic settings. More recently, LLMs like ChatGPT and GPT-4 that can answer a variety of user questions seem quite realistic. Because this content is well-written, LLMs are a convincing alternative to human labour in a variety of situations. This has raised worries, though, regarding the possibility that they may be abused to disseminate false information and impede the process of teaching and learning. To lessen the possibility of this kind of content being misused, automatic mechanisms for identifying machine-generated text must be developed, as humans only slightly outperform chance in this area. [10].

ChatGPT - The AI research company OpenAI created ChatGPT, or Chat Generative Pre-training Transformer, a sophisticated text generating tool. ChatGPT is a kind of language model that works well in conversational AI applications such as language learning programs, chatbots, and other applications that require human-like answers to input. [11].

Dolly-v2 - The big language model dolly-v2-12b from Databricks is licensed for commercial use and follows instructions. It was trained using the Databricks machine learning platform. Dolly receives her training from Databricks employees using approximately 15k instruction/response fine-tuning records (databricks-dolly-15k) in capability domains from the Instruct paper, such as brainstorming, classification, closed QA, generation, information extraction, open QA, and summarization. This is based on Pythia-12b. [7].

Cohere - Businesses may use their platform and tools to develop apps and services that take use of text production and natural language processing capabilities. The NLP models from Cohere are made to comprehend and produce text that is like that of a human, which makes it simpler for developers to build chatbots, virtual assistants, content creation tools, and other applications. Cohere offers developers APIs (Application Programming Interfaces) to include their text generating capabilities—which are based on sophisticated machine learning techniques—into their apps. This enables companies to leverage natural language creation and understanding to give their users interactive and engaging experiences. [7].

DaVinci – It is the GPT-3 language model from OpenAI that is available for purchase. The cutting-edge language model known as GPT-3, or "Generative Pre-trained Transformer 3," is intended to produce writing that appears human depending on the input it is given. Among the many models created by OpenAI with the GPT-3 architecture is DaVinci. Text creation tasks (including natural language comprehension, text completion, language translation, content production, and much more) are among the many tasks for which the DaVinci model is renowned. It can produce content in a variety of languages and disciplines that is coherent and pertinent to the situation.

Literature Survey

Recent advances in LLM can be traced back to those reported by Vaswani et al. Proposed transformer architecture. [3] introduces a self-attention mechanism that allows the model to focus on different regions of the input sequence. Later, Radford et al. [4] develop a generative pre-trained transformer (GPT) based on the transformer architecture. Because GPT is trained on a large corpus of text data, it achieves good performance on a variety of language generation tasks. GPT2 [5], a larger version of GPT that includes more parameters and is trained on a larger corpus, was developed and achieves better performance than GPT. GPT3 [6] is the third generation of GPT with over 175 billion parameters and is proven to produce consistent and context-appropriate text even in situations where minimal input or guidance is required. I am. Since November 2022, OpenAI has released his ChatGPT [7]. It is trained on his GPT 3.5 architecture and uses reinforcement learning from human feedback (RLHF) [8, 9] to improve its generative capabilities. Innovative text generation skills are displayed by ChatGPT, which can provide consistent and pertinent content for a range of applications, including chatbots, customer support, and education. Xin lei He and colleagues' work provides a thorough examination of the approaches currently in use for identifying machine-generated text (MGT) when strong Language Models (LLMs) are at play. They examine six metric-based and four model-based detection techniques, testing their efficacy on a variety of datasets. Among all the approaches, the LM Detector is the most successful and resilient in identifying MGTs with fewer words or those produced by distinct LLMs.

The researchers next investigate how to apply MGT identification techniques to the trickier job of text attribution. In this case, model-based techniques—particularly the LM Detector—outperform metric-based techniques because they are more skilled at capturing the syntactic and semantic connections between words and sentences. Although metric-based approaches have difficulty differentiating source LLMs, all approaches show potential for improvement in precisely identifying the source of MGTs.

Through the addition of adversarial perturbations to MGTs, the study explores the resilience of MGT detection techniques even more. The ChatGPT Detector exhibits notable susceptibility to minor disturbances, underscoring the urgent necessity for the advancement of more robust MGT detection techniques.

The researchers present MGTBench, a modular framework that integrates datasets and detection techniques, as a forward-thinking endeavour. This novel technology is well-positioned to act as a standard, enabling further studies to improve MGT detection techniques and enhance LLM training protocols. The study emphasises how important it is to keep working to improve the reliability and effectiveness of MGT detection techniques.

Methodology

The second model we tested was the bag of words model with frequency inverse literature terminology frequency point. Here the documents also are represented as a vector but instead of "0" and `1`, the document now contains points for each form. These scores are calculated by multiplying TF and IDF for specific words. So, the point of any word in any document can be represented as follows equation:

$$TFIDF(word,doc)=TF(word,doc)* IDF(word)$$

Two matrices are calculated in this method: one containing the inverse document frequency of a word in the entire corpus, and the other containing the frequency term of each word in each document. The formulas for calculating both matrices are as follows:

$$TF\ (word, doc) = Frequency\ of\ word \in thedoc/No.\ ofwords \in thedoc$$

$$IDF(word) = loge\ (1 + No.\ ofdocs/No.\ of\ docs\ withword)$$

Using this method, the provided example sentences can be converted into a TF-IDF model. An IDF dictionary is generated, containing frequent words and their corresponding IDF values. Additionally, a TF dictionary is built with given TF values for each word in each document [10].

This TF-IDF model differs from the binary bag-of-words model as it assigns more precise values between 0 and 1 instead of representing documents as vectors with "0" and "1". While it performs well overall and assigns greater importance to uncommon words compared to a binary bag-of-words model, it may struggle with sentences containing negation. Negation, a common linguistic structure influencing word polarity, suggests the need for a model that considers the presence of negatives for better performance, especially in grammatically complex scenarios [10].

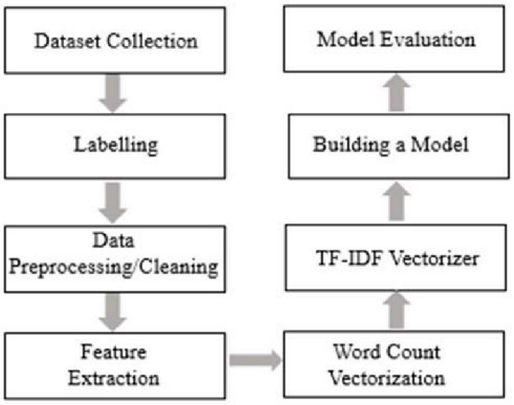
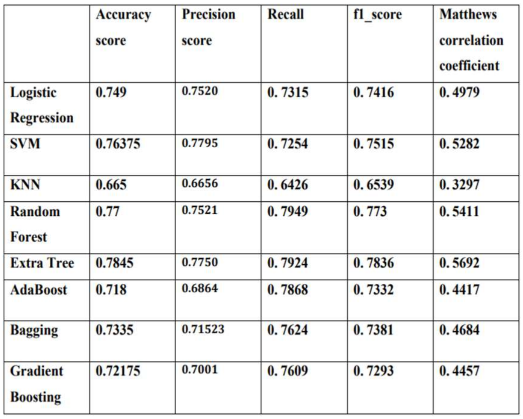


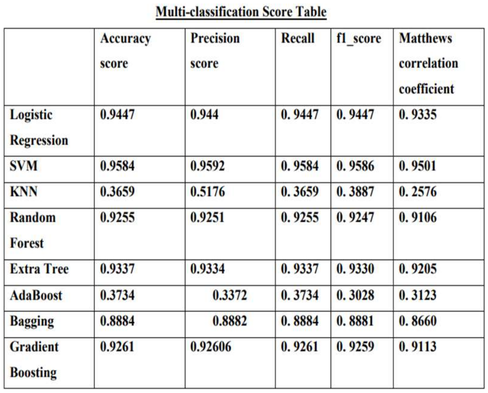
Fig. 1

Result



	Accuracy score	Precision score	Recall	f1_score	Matthews correlation coefficient
Logistic Regression	0.749	0.7520	0. 7315	0. 7416	0. 4979
SVM	0.76375	0.7795	0. 7254	0. 7515	0. 5282
KNN	0.665	0.6656	0. 6426	0. 6539	0. 3297
Random Forest	0.77	0.7521	0. 7949	0. 773	0. 5411
Extra Tree	0.7845	0.7750	0. 7924	0. 7836	0. 5692
AdaBoost	0.718	0.6864	0. 7868	0. 7332	0. 4417
Bagging	0.7335	0.71523	0. 7624	0. 7381	0. 4684
Gradient Boosting	0.72175	0.7001	0. 7609	0. 7293	0. 4457

Table 1



Multi-classification Score Table

	Accuracy score	Precision score	Recall	f1_score	Matthews correlation coefficient
Logistic Regression	0.9447	0.944	0. 9447	0. 9447	0. 9335
SVM	0.9584	0.9592	0. 9584	0. 9586	0. 9501
KNN	0.3659	0.5176	0. 3659	0. 3887	0. 2576
Random Forest	0.9255	0.9251	0. 9255	0. 9247	0. 9106
Extra Tree	0.9337	0.9334	0. 9337	0. 9330	0. 9205
AdaBoost	0.3734	0.3372	0. 3734	0. 3028	0. 3123
Bagging	0.8884	0.8882	0. 8884	0. 8881	0. 8660
Gradient Boosting	0.9261	0.92606	0. 9261	0. 9259	0. 9113

Table 2

Accuracy:

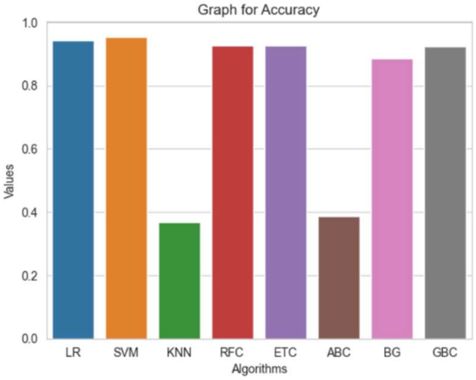


Fig. 2

Precision Score:

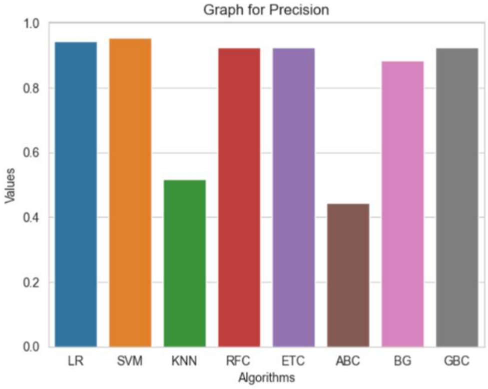


Fig. 3

Recall Scores:

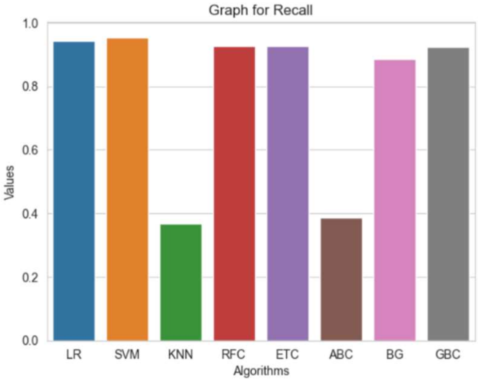


Fig. 4

3.3.4. F1 Scores:

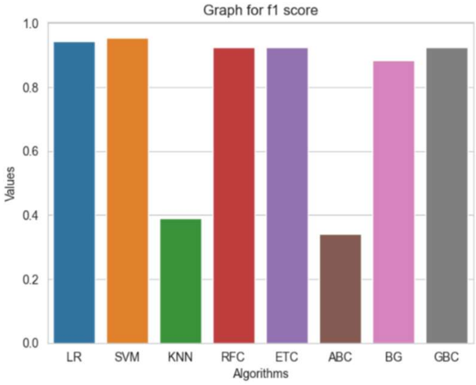


Fig. 5

Matthews Correlation Coefficient (MCC) Scores:

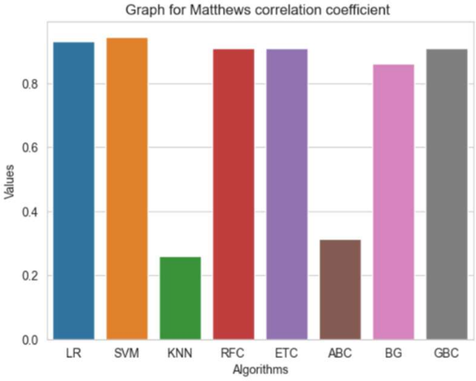


Fig. 6

References

[1] A. Harada, D. Bollegala and N. P. Chandrasiri, "Discrimination of human-written and human and machine written sentences using text consistency," 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 2021, pp. 41-47, doi: 10.1109/ICCCIS51004.2021.9397237.

[2] H. Alamleh, A. A. S. AlQahtani and A. ElSaid, "Distinguishing Human-Written and ChatGPT-Generated Text Using Machine Learning," 2023 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 2023, pp. 154-158, doi: 10.1109/SIEDS58326.2023.10137767.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In Annual Conference on Neural Information Processing Systems (NIPS), pages 5998–6008. NIPS, 2017.

[4] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. 2016.

[5] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. OpenAI blog, 2019.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel HerbertVoss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In Annual Conference on Neural Information Processing Systems (NeurIPS). NeurIPS, 2020.

[7] ChatGPT. https://chat.openai.com/chat.

[8] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep Reinforcement Learning from Human Preferences. In Annual Conference on Neural Information Processing Systems (NIPS), pages 4299–4307. NIPS, 2017.

[9] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize from human feedback. CoRR abs/2009.01325, 2020.

[10] Das, B., & Chakraborty, S. (2018). An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation. *ArXiv, abs/1806.06407*.