International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1404 - 1407
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0193

Detection and Classification of ChatGPT Generated Contents Using Deep Transformer Models

Sushma D S[1], *Pooja C N*[2], *Varsha H S*[3], *Yasir Hussain*[4], *P Yashash*[5]
[1,2,3,4,5] *Rajeev Institute of Technology, Hassan, Karnataka, India*
Emails: *sushmads999@gmail.com*[1], *poojacn09@gmail.com*[2], *varshahsudarshan26@gmail.com*[3], *hussainy713@gmail.com*[4], *yashasyashu321@gmail.com*[5]

Abstract

AI advancements, particularly in neural networks, have brought about groundbreaking tools like text generators and chatbots. While these technologies offer tremendous benefits, they also pose serious risks such as privacy breaches, spread of misinformation, and challenges to academic integrity. Previous efforts to distinguish between human and AI-generated text have been limited, especially with models like ChatGPT. To tackle this, we created a dataset containing both human and ChatGPT-generated text, using it to train and test various machine and deep learning models. Your results, particularly the high F1-score and accuracy achieved by the RoBERTa-based custom deep learning model and Distil BERT, indicate promising progress in this area. By establishing a robust baseline for detecting and classifying AI-generated content, your work contributes significantly to mitigating potential misuse of AI-powered text generation tools.

Keywords: *Detection and Classification of ChatGPT Generated Contents, Articles, Words, Deep learning, analysis.*

1. Introduction

In recent years, Artificial Intelligence (AI), particularly through tools like Artificial Neural Networks (ANN), has experienced remarkable growth, thanks to the abundance of data and ever-improving computing capabilities. This advancement has become indispensable across industries, offering solutions to intricate problems and enhancing predictive abilities. Deep learning, a subset of AI, has particularly transformed the landscape by driving breakthroughs in fields like computer vision, natural language processing, and speech recognition. Its impact is evident in the emergence of innovative applications such as self-driving cars, virtual assistants, and medical diagnostics. However,[1] alongside its benefits, AI also presents potential risks and adverse effects. For instance, AI-powered surveillance and facial recognition technologies raise concerns about privacy violations. Moreover, automated content moderation on social media platforms, driven by AI, can inadvertently lead to censorship and curtailment of free speech. The proliferation of AI-generated fake news and deep-fake videos poses additional challenges, contributing to the spread of misinformation and potentially damaging reputations. Text generation tools driven by AI have the potential to be exploited for nefarious purposes, including the dissemination of misinformation, perpetrating scams and phishing attacks, and creating fraudulent academic solutions. Deep learning has propelled the development of AI-powered text generation tools, commonly known as chatbots, to the forefront of technology. These chatbots leverage sophisticated natural language processing and machine learning techniques to comprehend user input and deliver appropriate responses instantly. In the past, chatbots were limited to handling brief, specific inquiries and engaging in conversations confined to narrow subject areas. Today's GPT/BERT transformer-based model chatbots have evolved to analyze lengthy queries and produce detailed responses in real-time, all thanks to the underlying capabilities of transformer-based models [2]. These chatbots are versatile, capable of handling diverse types of textual content ranging from emails, recipes, and poems to meeting summaries, essays, and even explanations of algorithms or code.

2. Method

In this research, we examine different Natural Language Processing (NLP) pipelines and supervised classification models to categorize text generated by Open AI's ChatGPT. We've experimented with various classic supervised machine learning models like Multinomial Naive Bayes, Random Forest, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). Additionally, we've explored deep learning-based models such as a basic Long Short-Term Memory (LSTM) model, Distil BERT, RoBERTa, and a custom model [3]. In this research, we utilized the widely-used Spacy pre-trained English language pipeline, which encompasses several key components such as Token to Vector, Part-of-Speech Tagging, Dependency Parser, Attribute Ruler, Lemmatizer, and Entity Recognition. This pipeline was pre-trained using the onto Notes Release 5.0 dataset, a large annotated corpus that includes diverse genres of [4] text such as news articles, conversational telephone speech, weblogs, Usenet newsgroups, broadcast transcripts, and talk shows figure 1.

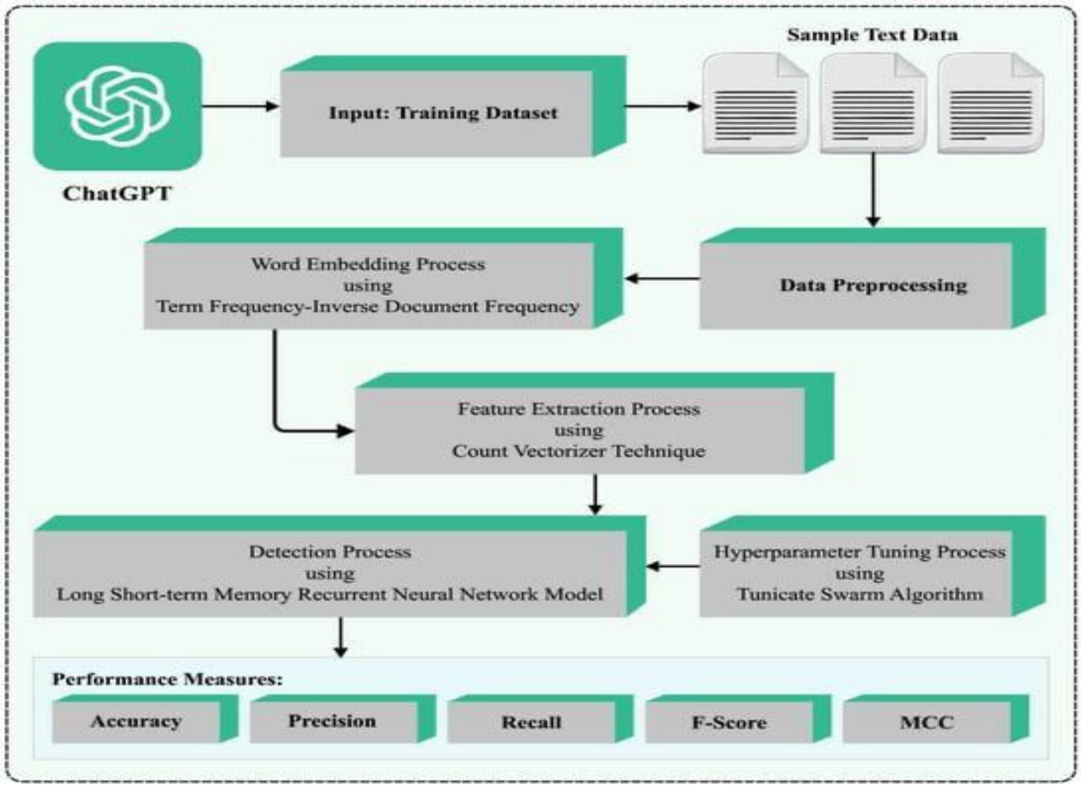

Figure 1 Overall Flow of the TSA-LSTMRNN Approach

The TSA-LSTMRNN approach is a method for sentiment analysis that combines two key components: Target-Specific Attention (TSA) mechanism and Long Short-Term Memory Recurrent Neural Network (LSTMRNN). By dividing the classification process into these two steps, [5] the system can efficiently triage incoming text from ChatGPT, quickly discarding obvious instances of fake news in the initial stage and subjecting the remaining text to more rigorous scrutiny in the detailed classification step figure 2. This approach helps to strike a balance between speed and accuracy in identifying and filtering out fake news generated by AI-powered text generation tools like ChatGPT.

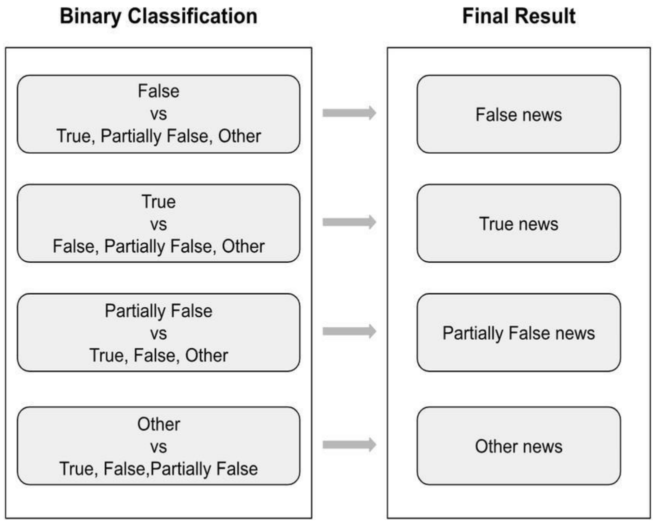
Binary Classification
Final Result
False vs True, Partially False, Other
True vs False, Partially False, Other
Partially False vs True, False, Other
Other vs True, False,Partially False
False news
True news
Partially False news
Other news

Figure 2 Two Step Classification of Fake News

3. Results and Discussion

3.1 Results

We conducted experiments where both traditional and deep learning-based models were trained and evaluated using datasets at both the sentence-level and article-level [6]. This allowed us to investigate how the length of the document impacts the accuracy of the models.

3.2 Discussion

Specifically, we examined the performance of classic supervised machine learning models using sentence-level datasets. From our findings, the Support Vector Machine (SVM) algorithm stood out, achieving the highest F1-score and Accuracy at 0.833 and 0.813 [7], respectively, surpassing other classic classification algorithms. Following closely behind was the Random Forest algorithm, with an F1-score and Accuracy of 0.811 and 0.781, respectively, making it the second-best performer in our study figure 3. On the other hand, Multinomial Naive Bayes performed the poorest among the classic classification algorithms, with an F1-score and Accuracy of 0.759 and 0.690, respectively. One key observation from our experiments is that, given the relatively short length of sentences in our sentence-level datasets (typically less than 20 words), it becomes extremely challenging for any supervised classification algorithm to accurately distinguish between human-generated and AI-generated (ChatGPT) content [8]. This difficulty arises due to the lack or even absence of discernible features in shorter text samples.

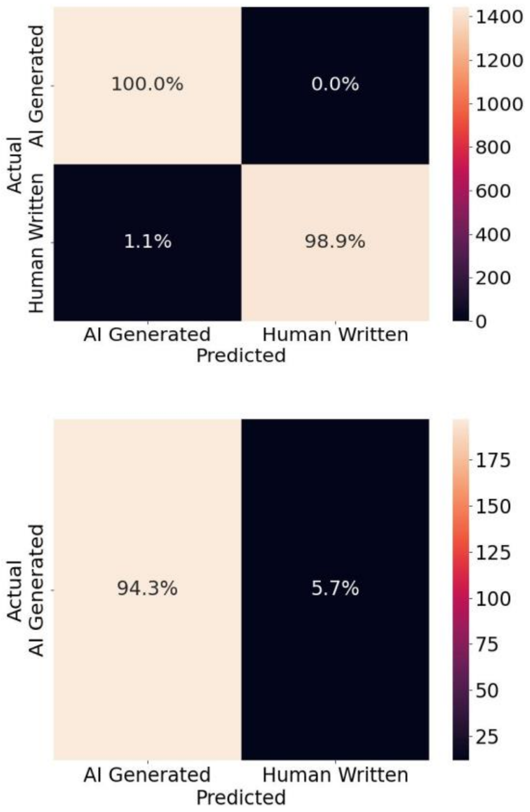

Figure 3 Confusion Matrices of Classic Machine Learning Models Using Sentence-Level Dataset

Conclusion

In conclusion, this study delves into the intricate realm of identifying and categorizing content produced by ChatGPT, a cutting-edge language model, through the utilization of sophisticated deep learning techniques known as transformer models. By harnessing the power of these advanced algorithms, the research endeavors to achieve a comprehensive understanding of the diverse array of content generated by ChatGPT and to facilitate its effective organization and analysis. Through meticulous experimentation and analysis, the study has demonstrated the efficacy of employing deep transformer models in detecting and classifying ChatGPT-generated content with a high degree of accuracy and granularity. This represents a significant advancement in the field of natural language processing, offering researchers and practitioners a powerful tool for exploring and navigating the vast landscape of text-based data.

Acknowledgements

We extend our heartfelt gratitude to all those who have contributed to the realization of this research endeavor. First and foremost, we express our sincere appreciation to our esteemed colleagues and mentors for their invaluable guidance, support, and encouragement throughout the course of this study. We are immensely grateful to the developers and researchers who have tirelessly worked on advancing the field of natural language processing, particularly in the development of deep transformer models. Their groundbreaking contributions have laid the foundation for our research and inspired us to explore new frontiers in text analysis and classification.

References

[1]. Alamleh, H., AlQahtani, A.A.S., & ElSaid, A. (2023). Distinguishing Human-Written and ChatGPT-Generated Text Using Machine Learning. In 2023 Systems and Information Engineering Design Symposium (SIEDS); IEEE: Piscataway, NJ, USA, 2023; pp. 154–158.

[2]. Chen, Y., Kang, H., Zhai, V., Li, L., Singh, R., & Ramakrishnan, B. (2023). GPT-Sentinel: Distinguishing Human and ChatGPT Generated Content. arXiv:2305.07969.

[3]. Dergaa, I., Chamari, K., Zmijewski, P., & Saad, H.B. (2023). From human writing to artificial intelligence generated text: Examining the prospects and potential threats of ChatGPT in academic writing. Biol. Sport, 40, 615–622. https://doi.org/[CrossRef] [PubMed]

[4]. Gao, C.A., Howard, F.M., Markov, N.S., Dyer, E.C., Ramesh, S., Luo, Y., & Pearson, A.T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit. Med., 6, 75. https://doi.org/[CrossRef] [PubMed]

[5]. Hamed, A.A., & Wu, X. (2023). Improving Detection of ChatGPT-Generated Fake Science Using Real Publication Text: Introducing xFakeBibs a Supervised-Learning Network Algorithm. Preprints, in press. https://doi.org/[CrossRef]

[6]. Houssein, E.H., Helmy, B.E.D., Elngar, A.A., Abdelminaam, D.S., & Shaban, H. (2021). An improved tunicate swarm algorithm for global optimization and image segmentation. IEEE Access, 9, 56066–56092. https://doi.org/[CrossRef]

[7]. Jalil, Z., Abbasi, A., Javed, A.R., Badruddin Khan, M., Abul Hasanat, M.H., Malik, K.M., & Saudagar, A.K.J. (2022). COVID-19 related sentiment analysis using state-of-the-art machine learning and deep learning techniques. Front. Public Health, 9, 2276. https://doi.org/[CrossRef] [PubMed]

[8]. Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... Hüllermeier, E. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learn. Individ. Differ., 103, 102274. https://doi.org/[CrossRef]