Preprint accepted at the 14th International Conference on Computer and Knowledge Engineering (ICCKE 2024). Find the published version at https://ieeexplore.ieee.org/document/10874567

Distinguishing Abstracts of Human-Written and ChatGPT-Generated Papers in the Field of Computer Science

Mohsen Arzani
PerLab, Faculty of electrical and computer engineering
University of Birjand
Birjand, Iran
mohsenarzani@birjand.ac.ir

Hamed Vahdat-Nejad*
PerLab, Faculty of electrical and computer engineering
University of Birjand
Birjand, Iran
vahdatnejad@birjand.ac.ir

Matin Hossein-Pour
PerLab, Faculty of electrical and computer engineering
University of Birjand
Birjand, Iran
matin.hosseinpour@birjand.ac.ir

Abstract— ***In the era of artificial intelligence, advanced technologies such as text deep fakes have emerged, utilizing AI and deep learning algorithms to generate text contents that convincingly mimic reality. Text-based deep fakes, commonly found in emails, articles, news, and social media posts, pose a significant threat to public trust. Our research proposes a method to distinguish between fake and genuine scientific abstracts in the field of computer science using a specialized dataset and deep learning models. The proposed detection model can differentiate between real and fake abstracts with 97.5% accuracy. Additionally, we employed metrics such as precision, recall, accuracy, and F1 score to measure the performance of our system in detail.***

Keywords— *LLM-generated text, Deepfake text, Deep learning, ChatGPT*

I. INTRODUCTION

In today's digital world, rapid advancements in artificial intelligence (AI), have led to the emergence of technologies such as deep fake [1]. These technologies utilize AI and deep learning algorithms to generate video, audio, image, and text content that convincingly resembles reality [2]. These contents have a profound impact on the digital society. Deep fakes can create images and videos where individuals' faces are seamlessly replaced [3], or generate speech and voices that seem to come from famous and credible individuals [4]. While this technology can be used for positive purposes, such as creating educational and entertainment content, it also poses a serious threat to public trust in digital information and media.

Among the various types of deep fakes, text-based deep fakes represent an advanced application of AI and natural language processing, producing fake yet believable written content [5]. This technology employs large language models [6], such as Chat GPT [7], capable of analyzing and generating highly accurate and refined texts. Text-based deepfakes can appear in emails, articles, news, and social media posts [8], mimicking human writing styles to the point where superficial detection becomes very difficult. If people are unable to properly verify the accuracy of information, their trust in media and digital platforms diminishes. This issue can undermine democracy, promote conspiracy theories, and increase social instability [9].

The methods for detecting text deepfakes are divided into manual and automatic approaches. However, there has been little research focused on detecting fake scientific abstracts specifically in the field of computer science. Existing studies have addressed broader categories of fake texts without a particular focus on scientific literature, especially within computer science.

Our proposed approach addresses this gap by detecting fake scientific abstracts in the field of computer science generated by ChatGPT. We have specifically collected a dataset of abstracts from computer science papers and developed a deep learning model using advanced techniques such as BERT and RoBERTa. This approach allows us to accurately distinguish between human-written abstracts and those generated by ChatGPT.

To evaluate our proposed model, we use metrics such as accuracy, recall, precision, and F1 score. The results show that the RoBERTa model performs exceptionally well, achieving an accuracy of 97.5%.

Preprint accepted at the 14th International Conference on Computer and Knowledge Engineering (ICCKE 2024). Find the published version at https://ieeexplore.ieee.org/document/10874567

Our findings demonstrate that the proposed approach is highly effective in detecting fake scientific abstracts in the field of computer science. This research highlights the potential of advanced AI models for detecting text deepfakes.

In the second section, we discuss related research. The third section describes our proposed method. The fourth section presents the experiments and evaluation results, and the final section concludes the paper.

II. RELATED WORK

Despite the various efforts made in the field of fake text detection across different domains, comprehensive research on detecting fake abstracts has not yet been conducted, making this area a promising field for novel research. Our research focuses on detecting abstracts of scientific papers in the field of computer science, which appears to be an area with limited significant research.

Among the published papers, there are only two similar studies in this domain, but neither of these studies has sufficient focus and precision for detecting computer science papers and they cannot perform effectively. In 2023, researchers at the University of Thessaly in Greece focused on detecting fake abstracts in the context of COVID-19 [10]. They used a dataset comprising 28,662 scientific abstracts, half generated by Chat GPT and the other half selected from articles in the CORD-19 dataset. By combining text representation techniques with machine learning methods, they achieved a detection accuracy of 98.7%.

Furthermore, in 2024, additional research on detecting fake text in scientific papers was conducted [11]. This time, a system named AI-Catcher was developed, based on a combination of MLP and CNN features. The system achieved an excellent detection accuracy of 98.8%.

Although these studies have achieved high accuracy in detecting fake texts generated by Chat GPT, they are not as effective for computer science papers. It seems that with more focus on this area and the use of domain-specific datasets, better results and improved detection accuracy can be achieved.

III. RESEARCH METHODOLOGY

In this section, we describe the proposed method. The methodology uses machine learning to systematically detect the abstracts generated by CHAT GPT.

A. Dataset

Since this research requires a dataset of abstracts from scientific articles in the field of computer science and such a dataset is not readily available, we need to create a dataset comprising human-written abstracts and those generated by Chat GPT. To address this need, we first extract the computer science subcategories from the ACM taxonomy. Then, for each category, we select a reputable journal from each of the four leading computer science publishers: ACM, IEEE, Elsevier, and Springer. From each journal, we select 20 articles published between 2021 to 2024.

For each selected article, we collect and record information including the article's title, abstract, journal, publisher, year of publication, category in computer science, DOI of the article, and the journal's impact factor. Subsequently, using the following prompt, Chat GPT generates corresponding abstracts, which are also added to the dataset:

Give me the abstract of a new article entitled "Article Title" to be submitted to the "Name of the journal ".

Ultimately, we have a dataset containing 2080 computer science article abstracts, consisting of 1040 published abstracts and 1040 corresponding abstracts generated by Chat GPT.

B. Data preprocessing

Data preprocessing is one of the critical steps in preparing text for processing. In this study, the preprocessing steps are as follows:

- Stopword Removal: Stopwords are words that do not carry significant meaning on their own and are repeated in most texts, such as "and," "of," "to," etc [12]. These words can create noise and should be removed. Removing stopwords helps us focus more on important words and key concepts in the text, which can improve the accuracy of machine learning models.

- Stemming: This step involves converting words to their root forms. Stemming helps to standardize different words with similar meanings, such as converting "books," "my books," and "their books" to "book." By using stemming, the number of text features is reduced, which can help decrease the complexity of models and increase their efficiency.

-Removing Whitespace and Punctuation: Extra whitespace and punctuation can create noise and should be removed to prepare the text for modeling in a cleaner form. This step includes removing extra spaces, commas, periods, and other unnecessary punctuation marks. Doing this results in more cohesive and consistent data, which can improve model performance.

Proper data preprocessing is one of the keys to success in natural language processing and machine learning. These steps help prepare the data in an optimal way for modeling, improving the accuracy and efficiency of the models.

C. Model Description

We utilize advanced machine learning models such as BERT [13] and RoBERTa [14], which are renowned for their high accuracy and performance in natural language processing. These models have been extensively used in various research papers and studies, demonstrating their effectiveness in analyzing and understanding natural language texts. BERT and RoBERTa models leverage deep transformer neural network architectures, which excel in comprehending the complex dependencies and relationships between words and phrases.

Preprint accepted at the 14th International Conference on Computer and Knowledge Engineering (ICCKE 2024). Find the published version at https://ieeexplore.ieee.org/document/10874567

Initially, these models were trained on large and diverse datasets such as Wikipedia and Book Corpus to acquire extensive general language knowledge. However, to enhance their performance for specific applications tailored to our research needs, these models are retrained with our specific dataset (a process known as fine-tuning). This process optimizes the models to recognize the particular patterns and concepts of interest, resulting in higher accuracy and efficiency in specific tasks.

To evaluate the accuracy of the proposed system, we use various metrics. Precision indicates the accuracy of the model's positive predictions. Accuracy reflects the ratio of correct predictions to the total number of samples. Recall measures the model's ability to identify all positive instances in the data, and the F1-score combines precision and recall, balancing the two metrics.

These metrics help us comprehensively assess the performance of our proposed models and make necessary improvements as needed. By using this approach, we can develop an advanced natural language processing system capable of precise and effective detection and analysis.

IV. EXPERIMENTAL RESULTS

In this section, we present the results obtained from BERT and RoBERTa machine learning models for classification. Our task is to detect and classify fake versus genuine abstracts of computer science papers. The training set includes 80% of the dataset, consisting of 1664 abstracts, and the test set includes 20% of the dataset, consisting of 416 abstracts.

As mentioned, in our proposed method, we first preprocess the abstracts using various techniques such as stop-word removal, stemming, and the elimination of whitespace and punctuation. This ensures that cleaner and better-formatted text is fed into the models. We then used two machine learning models, BERT and RoBERTa, for classification. Although the BERT model performed well with an accuracy and F1 score of 96.8%, the RoBERTa model achieved even better results with an accuracy of 97.5%. Considering the inherent features of RoBERTa, this model has shown better performance compared to BERT due to its longer pre-training phase, the removal of the next sentence prediction task, and its superior memory for longer texts. These characteristics have helped our model detect more subtle differences between human-generated scientific abstracts and those generated by ChatGPT.

Table 1 presents the evaluation results using the metrics of Precision, Accuracy, Recall, and F1-score for both classes: Human and Chat GPT.

TABLE I. RESULTS OF EXPERIMENTS

Model	Human			CHAT GPT			All
	Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*	*Accuracy*
Bert	0.976	0.962	0.969	0.961	0.975	0.968	0.968
RoBERTa	0.995	0.957	0.976	0.957	0.995	0.975	0.975

As seen in Table 1, the higher accuracy belongs to the RoBERTa model. In the RoBERTa model, the scores for the Precision, Recall, and F1-score metrics for the class of abstracts written by humans are 0.995, 0.957, and 0.976, respectively. Similarly, for the class of abstracts written by Chat GPT, the scores are 0.957, 0.995, and 0.975, respectively.

V. CONCLUSION

By focusing on detecting scientific abstracts generated by Chat GPT in the field of computer science, we developed an efficient and robust method using advanced machine learning models like BERT and RoBERTa. Our results indicate that when these models were trained with our specialized dataset, they were able to distinguish between real and fake abstracts with high precision, accuracy, recall, and F1 score. Specifically, the RoBERTa model demonstrated impressive performance with an accuracy of 97.5%, proving its potential in this domain.

The significance of our research is considerable, as it lays the foundation for future studies and the development of more advanced detection systems. By effectively identifying and mitigating the risks associated with text-based deep fakes, we can maintain public trust in digital information and enhance the credibility of scientific research. Future works may focus on expanding the dataset to encompass a broader range of scientific fields and utilizing additional features or models to enhance detection capabilities.

VI. REFERENCES

[1] J. W. Seow, M. K. Lim, R. C. Phan and J. K. Liu, "A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities," *Neurocomputing*, vol. 513, pp. 351-371, 2022.

[2] T. Zhang, "Deepfake generation and detection, a survey," *Multimedia Tools and Applications*, vol. 81, no. 5, pp. 6259-6276, 2022.

[3] M. S. Rana, M. N. Nobi, B. Murali and A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE access*, vol. 10, pp. 25494-25513, 2022.

[4] H. Cheng, Y. Guo, T. Wang, Q. Li, X. Chang and L. Nie, "Voice-face homogeneity tells deepfake," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 3, pp. 1-22, 2023.

[5] R. Mubarak, T. Alsboui, O. Alshaikh, I. Inuwa-Dutse, S. Khan and S. Parkinson, "A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats," *IEEE Access*, vol. 11, 2023.

[6] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan and M. Shah, "Transformers in Vision: A Survey," *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1-41, 2022.

Preprint accepted at the 14th International Conference on Computer and Knowledge Engineering (ICCKE 2024). Find the published version at https://ieeexplore.ieee.org/document/10874567

[7] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel and J. Pfeffer, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, 2023.

[8] S. M. Saravani, I. Ray and I. Ray, "Automated Identification of Social Media Bots Using Deepfake Text Detection," in *International Conference on Information Systems Security*, Patna, India, 2021.

[9] A. T. Y. Chong, H. N. Chua, M. B. Jasser and R. T. Wong, "Bot or Human? Detection of DeepFake Text with Semantic, Emoji, Sentiment and Linguistic Features," in *IEEE 13th International Conference on System Engineering and Technology*, Shah Alam, Malaysia, 2023.

[10] P. C. Theocharopoulos, P. Anagnostou, A. Tsoukala, S. V. Georgakopoulos and S. K. Tasoulis, "Detection of fake generated scientific abstracts," in *IEEE Ninth International Conference on Big Data Computing Service and Applications*, Athens, Greece, 2023.

[11] B. Alhijawi, R. Jarrar, A. AbuAlRub and A. Bader, "Deep Learning Detection Method for Large Language Models-Generated Scientific Content," *arXiv preprint arXiv:2403.00828*, 2024.

[12] N. K. Cannannore , S. B. Syed and D. Andreas, "Comparative study between traditional machine learning and deep learning approaches for text classification," in *DocEng '18: ACM Symposium on Document Engineering*, Halifax Regional Municipality, 2018.

[13] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692v1*, 2019.