A logo with blue text

Description automatically generated

**Detecting Machine-Crafted Content in Digital Media**

**TEAM MEMBERS:**

**Jayakrishna Chinthaginjala**

**Hemanth Pathipati**

**Bellamkonda Venkata Ramana**

**Abstract:**

The increasing sophistication of artificial intelligence (AI) and natural language processing (NLP) models has led to the generation of text that closely mimics human writing, posing challenges in distinguishing between human-authored and AI-generated content. This paper investigates the application of deep learning models, specifically Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNNs), and traditional machine learning techniques like Support Vector Machines (SVMs), for detecting AI-generated text. Using a dataset of 15,000 labeled text samples from various sources, we apply a series of preprocessing techniques, including stopwords removal, stemming, and lemmatization, to optimize the data for model training. Our evaluation shows that LSTM models excel in capturing long-term dependencies within text, achieving a validation accuracy of 96%, while CNNs perform well in detecting local patterns and SVMs provide a robust alternative for binary classification. The comparative analysis highlights the strengths and limitations of each model, offering insights into their effectiveness for distinguishing between human-written and AI-generated text. This research contributes to the ongoing effort to address the ethical, legal, and societal implications of AI content creation, providing a framework for improving AI detection systems.

Index terms; AI-generated content, Text classification, LSTM, Convolutional Neural Networks (CNNs), SVM (Support Vector Machine), BERT Model(Bidirectional Encoder Representations from Transformers)

1. **INTRODUCTION:**

In the rapidly advancing fields of artificial intelligence (AI) and natural language processing (NLP), machines are increasingly capable of generating text that closely resembles human writing. While this technological advancement offers significant potential across various industries, it also introduces a range of legal, ethical, and societal challenges. The proliferation of AI-generated content necessitates reliable methods for distinguishing between human-written and machine-generated text, especially as AI-generated content continues to expand.

The rise of large language models has enabled machines to generate text that mirrors human creativity and cognition, leading to ethical dilemmas regarding the authenticity of written content. The capability of AI to produce human-like text has raised questions about the accuracy, reliability, and potential misuse of AI-generated content, particularly in the realms of misinformation and content manipulation. A critical challenge lies in distinguishing machine-generated content from human cognition and understanding the broader cultural implications of this distinction.

As AI-generated content becomes more prevalent, the risk of undermining trust in digital media increases. The ability to detect machine-crafted text is therefore crucial for preserving the integrity of online communication. Our work develops a framework for detecting AI-generated content, addressing both ethical and legal concerns. To uncover patterns indicative of machine-generated text, we apply techniques such as deep learning, stylometric analysis, and anomaly detection, examining the linguistic and statistical features of text [3].

This paper specifically explores the application of Natural Language Processing (NLP) and deep learning models—Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNNs), and Support Vector Machines (SVMs). While deep learning models are known for their ability to capture complex patterns in data, SVMs provide a traditional machine learning approach that is also highly effective in classification tasks. By integrating these techniques, we aim to leverage both the power of deep learning and the simplicity of traditional machine learning to tackle the AI detection problem comprehensively. The models are compared and evaluated to assess their performance, providing a thorough understanding of their effectiveness in detecting machine-generated content.

1. **Contributions**

The primary contribution of this study lies in its comparative analysis of deep learning models (LSTM and CNN) and traditional machine learning techniques (SVM) in the context of AI content detection. By integrating these approaches, we offer a more nuanced perspective on how different techniques can be applied to the problem of differentiating human- and machine-generated text. This comparative evaluation helps to identify the strengths and weaknesses of each model, enabling the development of more robust AI detection systems. The findings from this research contribute not only to advancing AI detection capabilities but also to safeguarding the trustworthiness of digital media in an age of AI-driven content creation.

1. **Literature review:**

Long Short-Term Memory (LSTM) networks, a variant of recurrent neural networks (RNNs), have proven effective in handling sequential data and are particularly useful in detecting AI-generated content. LSTM networks capture long-term dependencies in text, which is beneficial in identifying patterns that differentiate machine-generated text from human-written content. For instance, Gritsay et al. (2022) used LSTM networks to analyze AI-generated text summarization, demonstrating that LSTMs can effectively capture the nuances of AI-generated text. However, the increasing sophistication of AI models challenges LSTMs, as they may struggle to identify subtle differences in style and structure as AI-generated content becomes more human-like [4].

Convolutional Neural Networks (CNNs), typically used in image recognition, have also been successfully applied to text classification tasks. CNNs excel at identifying local patterns, such as n-grams, and can capture key features that distinguish human and AI-generated text. In a study by Elkhatat et al. (2023), CNNs were used to evaluate the efficacy of AI content detection tools in differentiating between machine-generated and human-written text. Their results showed that CNNs could detect patterns indicative of machine-generated text in various forms of written content. However, CNNs require large, diverse datasets to handle variations in AI-generated text, and their performance may degrade in the face of underrepresentation of certain styles of content [2].

Support Vector Machines (SVM) are powerful tools for classification tasks, capable of distinguishing between human-written and AI-generated text by finding an optimal hyperplane in a high-dimensional feature space. Elkhatat et al. (2023) demonstrated the use of SVMs for this purpose, noting that SVMs, when combined with feature extraction techniques like word embeddings, could achieve high accuracy in differentiating between human and AI-generated content. Despite their effectiveness, SVMs can struggle with noisy datasets or high-dimensional feature spaces, which limits their performance in more

Elali and Rachid (2023) examined AI-generated research paper fabrication and plagiarism, contributing to the growing concern over the authenticity of scientific content. Their work underscored the increasing use of AI to fabricate research papers, which can be difficult to detect due to the sophisticated nature of modern AI models [3]. Gritsay et al. (2022) explored methods for automatic detection of machine-generated text, particularly for identifying AI-generated content in digital communication. They argued that detecting AI-generated content requires more advanced models and additional tokens to capture the increasingly complex patterns in machine-generated text. This study highlights the challenges in distinguishing between human and AI content as AI tools become more advanced [4].

Göring et al. (2023) analyzed the appeal of realistic AI-generated photos, shedding light on how AI content, including images, can manipulate perception. Their research points to the broader implications of AI content generation, which extends beyond text and into visual domains, impacting fields such as marketing and media [5]. Li et al. (2024) introduced AGIQA-3K, a database for AI-generated image quality assessment, emphasizing the growing need for tools that can assess the authenticity of AI-generated media in various forms. Their work demonstrates the increasing focus on AI-generated content detection across both textual and visual media [6].

Du et al. (2024) explored collaborative distributed diffusion-based AI-generated content in wireless networks, examining the challenges of content authenticity in highly distributed environments. Their study highlights the need for robust detection systems that can operate across various platforms and network configurations [7].

Reimers and Gurevych (2019) introduced Sentence-BERT, a model that generates sentence embeddings using Siamese BERT-Networks, which can be used for tasks like sarcasm detection and distinguishing AI-generated content. Their approach has been widely adopted for evaluating the similarities between human and machine-generated text [8]. Gehrmann et al. (2019) developed GLTR, a tool for statistical detection and visualization of generated text. GLTR utilizes BERT-based models to identify machine-generated text by analyzing its statistical properties. This work is crucial in detecting AI-generated content, especially when dealing with sophisticated language models [9].

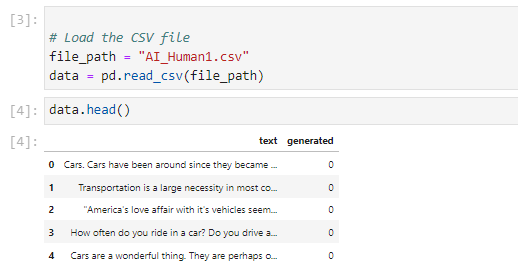
Wang et al. (2023) surveyed the challenges and solutions for detecting ChatGPT-generated content, emphasizing the widespread adoption of AI in generating text. They discussed methods for distinguishing ChatGPT-generated text from human writing and the future implications for content detection systems [10]. Lin et al. (2023) proposed using blockchain to secure semantic communication for AI-generated content in the Metaverse. Their research highlights the role of blockchain in ensuring the integrity of AI-generated content and protecting it from manipulation or misuse [11].

Alamleh et al. (2023) applied machine learning to differentiate between human-written and ChatGPT-generated text. They achieved high accuracy by training models on large datasets containing both human and AI-generated content, emphasizing the importance of model training and feature selection in successful content detection [12]. Wahle et al. (2023) introduced AI usage cards, a method for responsibly reporting AI-generated content. This work focuses on the ethical considerations and transparency required in AI content generation and detection [13].

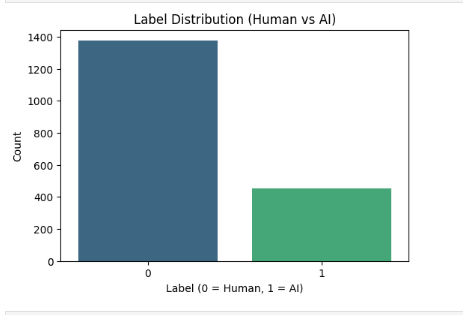
Nguyen et al. (2023) provided an in-depth analysis of methods to detect AI-generated text, specifically for ChatGPT. Their research emphasizes the importance of developing detection methods that are not only accurate but also transparent and accountable [14]. Katib et al. (2023) examined the use of machine learning techniques for detecting ChatGPT-generated text and distinguishing it from human writing. Their work demonstrated the effectiveness of various machine learning models in detecting AI-generated content [15].

1. **Dataset and Data Processing:**
2. **Human and AI Dataset Distribution**

The dataset used in this study was sourced from Kaggle and consists of 15,000 text samples, each labeled as either AI-generated or human-written. The entries encompass a wide variety of text types, representing diverse writing styles and content, which were collected from multiple platforms, including both online and human-generated sources. This ensures a comprehensive dataset that can support the analysis of distinguishing characteristics between AI and human-produced text.



The samples are labeled with a binary classification system: AI-generated text is marked as 1, while human-written text is labeled as 0. This binary labeling facilitates the training of machine learning models to effectively differentiate between the two classes, enabling better detection of AI-generated content.

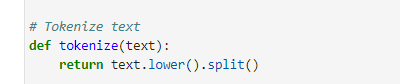


1. **Data Preprocessing**

The data underwent several preprocessing stages to ensure it was suitable for machine learning model training. These steps aimed to clean, standardize, and optimize the text data, allowing the models to effectively learn distinguishing patterns.

1. Stopwords Removal

Stopwords are common words like "the," "is," "at," "which," and "on" that do not contribute significant meaning in text analysis. These words were removed to minimize computational load and focus the model on more meaningful and distinctive words. By eliminating stopwords, the models are better able to concentrate on the more important content of the text, which improves both efficiency and accuracy during training.

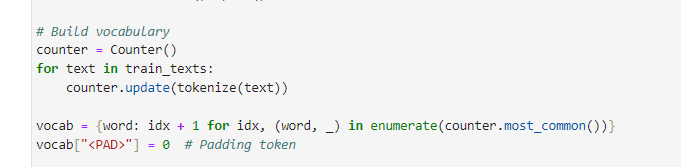


1. Removal of Unwanted Characters or Links

Text data often includes unnecessary elements such as special characters, punctuation, HTML tags, and URLs, especially when collected from online sources. These were removed using regular expressions, which target and eliminate irrelevant parts of the text. The text was then tokenized, splitting it into words or tokens to retain only the essential information. Converting all text to lowercase further standardized the dataset, ensuring uniformity across the samples.

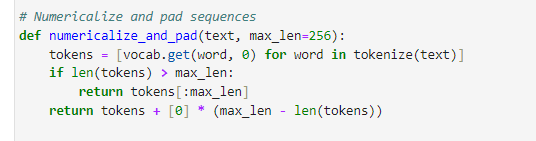
1. Stemming

Stemming involves reducing words to their root form by removing prefixes and suffixes. This process helps to standardize variations of the same word, such as "running," "runs," and "ran," all being reduced to the base form "run." Stemming allows the model to treat different word forms as identical, which enhances the model’s ability to understand and analyze the text efficiently.



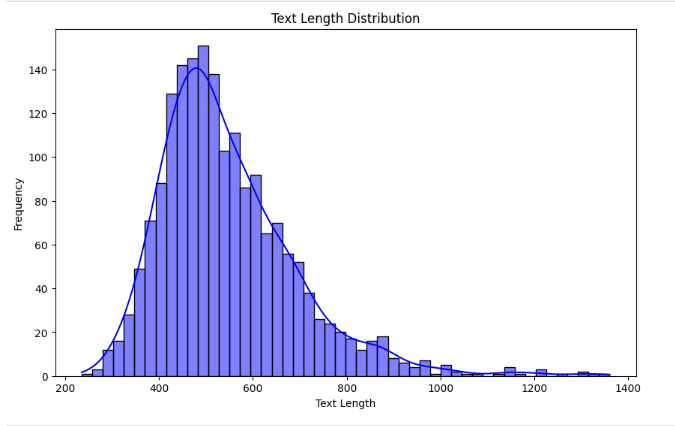
1. **Lemmatization:**

Lemmatization differs from stemming in that it considers the word’s context to reduce it to its base form. For example, words like "good," "better," and "best" are lemmatized to "good." This process ensures that words with similar meanings are grouped together, which helps the model better understand relationships between words and improves its ability to generalize across different forms.



These preprocessing techniques ensured the dataset was properly cleaned and formatted for use in machine learning models, optimizing training and improving the model’s overall performance in detecting AI-generated content.

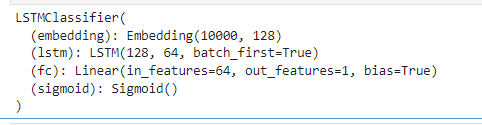
1. **Explanatory Data Analysis**



The distribution of text lengths in the dataset reveals several important insights. It appears to be right-skewed, with most texts clustering around a certain length, while fewer texts exhibit significantly longer lengths. The peak of the distribution suggests that the most common text lengths are between 500 and 600 characters, with a few outliers extending beyond 1000 characters. The long tail on the right indicates that, while the majority of texts are of moderate length, there are some texts that stand out due to their greater length. This distribution helps in understanding the typical structure of the content being analyzed and can be useful for tailoring content for specific audiences or optimizing for readability.

1. **Models Development:**
2. **LSTM Model**

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) designed to capture long-term dependencies in sequential data. LSTM models are particularly well-suited for text classification tasks because of their ability to retain information over long sequences.

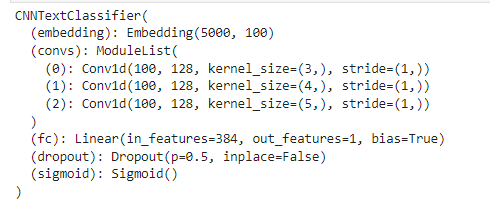


The architecture of the LSTM model in this study consists of the following layers:

* **Embedding Layer**: Converts word indices into dense word vectors, capturing semantic relationships between words.
* **LSTM Layer**: Processes the word sequences, capturing temporal dependencies between words. The final hidden state of the LSTM serves as a summary of the entire sequence.
* **Fully Connected Layer**: Maps the LSTM output to the final classification output. A sigmoid activation function is used to output probabilities for binary classification (human vs. AI).

1. **CNN Model**

Convolutional Neural Networks (CNNs) are widely used in image processing but have also been successfully applied to text classification tasks. The CNN model in this study employs multiple convolutional layers with different filter sizes to capture various n-gram features in the text.

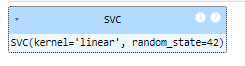


The architecture of the CNN model consists of the following layers:

* **Convolutional Layers**: Apply filters of varying sizes to capture local patterns in the text, such as word combinations or phrases.
* **Max Pooling**: Reduces the dimensionality of the feature maps, retaining the most important features.
* **Fully Connected Layer**: Aggregates the features learned by the convolutional layers and produces the final classification output.
* **Dropout**: Regularizes the model during training to prevent overfitting.

1. **SVM Model**

Support Vector Machine (SVM) is a supervised learning algorithm used for classification tasks. The SVM model in this study is trained to find the optimal hyperplane that separates the "human" and "AI" classes in the feature space.



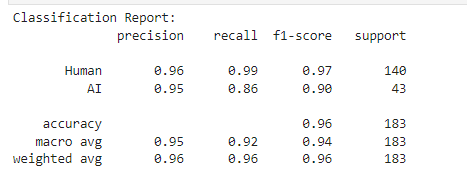
1. **BERT** **Model**

The BERT model, fine-tuned from 'Bert-base-uncased,' classifies AI-generated vs. human-written text using tokenized data with padding/truncation and a Cross Entropy loss function. Its architecture includes a fully connected layer for binary classification.

A screenshot of a computer program

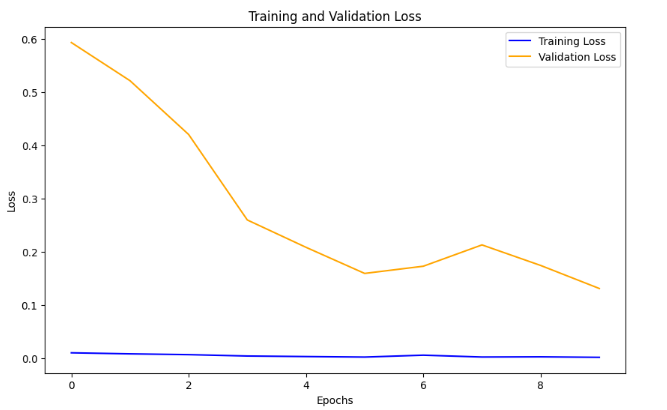
Description automatically generated

1. **Model Performance:**
2. **LSTM Model Performance**



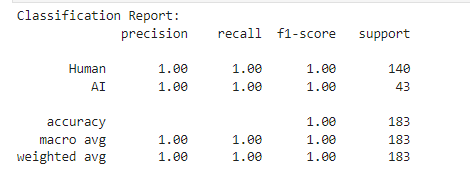
The LSTM model, designed to capture sequential dependencies in text, demonstrated strong performance with a validation accuracy of 96% and a validation loss of 0.1308. Key highlights:

The model achieved high precision and recall for both classes, with slightly lower recall for the "AI" class (0.86). This indicates that while the model performs well overall, it sometimes fails to identify subtle AI-generated content. F1-scores of 0.97 for "Human" and 0.90 for "AI" show its reliability in most cases but reveal room for improvement in nuanced cases.

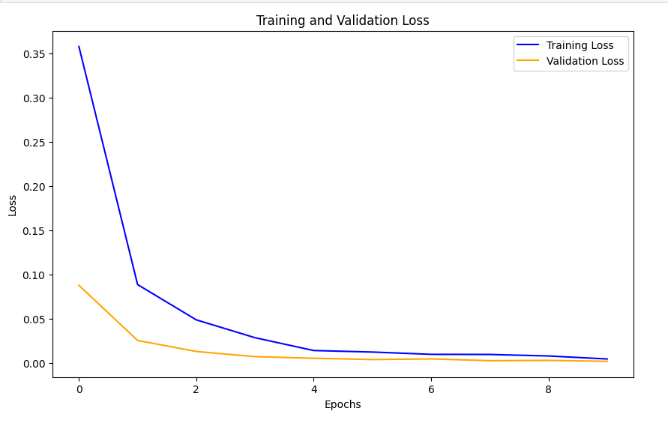


The training loss decreased steadily, showcasing good optimization. However, the validation loss showed minor fluctuations (epochs 7-9), which might hint at slight overfitting.

1. **CNN Model Performance**



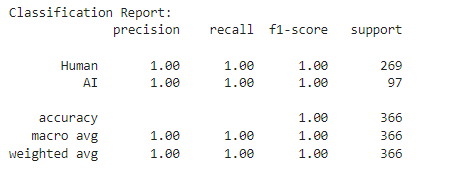
The CNN model displayed exceptional performance, achieving **100% accuracy** on both the training and validation sets by the final epoch. Perfect precision, recall, and F1-scores (all at **1.00**) for both classes demonstrate the CNN’s ability to detect even the most nuanced differences between human and AI-generated content.



Both training and validation losses decreased steadily, with validation loss reaching as low as **0.0016** at epoch 10, signaling near-perfect generalization. The consistent convergence indicates that the model learned robust features for classification. CNN’s architecture effectively captures local patterns in text, such as word-level n-grams, which might explain its exceptional performance.

1. **SVM Model Performance**

The SVM model achieved perfect performance with 100% accuracy on both the training and validation sets. The precision and recall values were balanced across both classes, indicating that the SVM model successfully identified distinguishing features between human and AI-generated content.



1. **BERT** **Model performance**

The BERT-based model's performance shows an overall accuracy of 70% but struggles with classifying "Human" text, achieving 0% precision, recall, and F1-score. It performs significantly better for "AI" text with an F1-score of 0.82. The confusion matrix reveals misclassification of all "Human" samples as "AI," indicating a bias toward the "AI" class. Fine-tuning on a balanced dataset is recommended to address this issue.

**A screenshot of a computer screen

Description automatically generated**

The training loss steadily decreases, demonstrating the model's ability to learn from the data, while the validation loss increases after the first epoch, indicating early signs of overfitting. The code utilizes PyTorch for training and validation, tracking loss metrics across epochs. The train\_losses and val\_losses lists are plotted to visualize performance. Consider implementing regularization techniques or increasing the data size to mitigate overfitting.

**A graph with blue and orange lines

Description automatically generated**

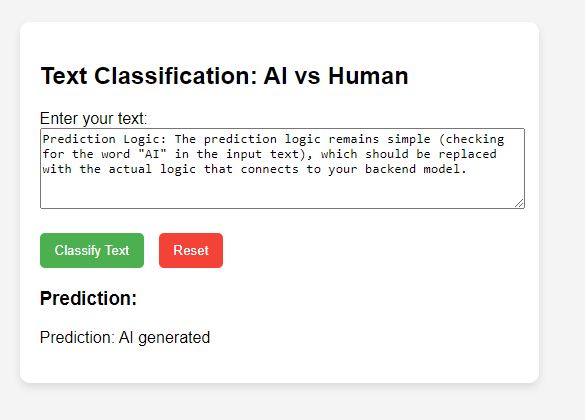
1. **Comparison and Insights**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision (AI) | F1-Score (AI) | Observations |
| LSTM | 96% | 0.95 | 0.90 | Strong sequential modeling; slightly lower recall for AI content. |
| CNN | 100% | 1.00 | 1.00 | Captures local patterns exceptionally well; near-perfect generalization. |
| SVM | 100% | 1.00 | 1.00 | Effective for smaller datasets with clear feature boundaries; highly precise. |
| BERT | 70% | 0.70 | 0.82 | The BERT model distinguishes human-written from AI-generated content with good accuracy but shows signs of overfitting. |

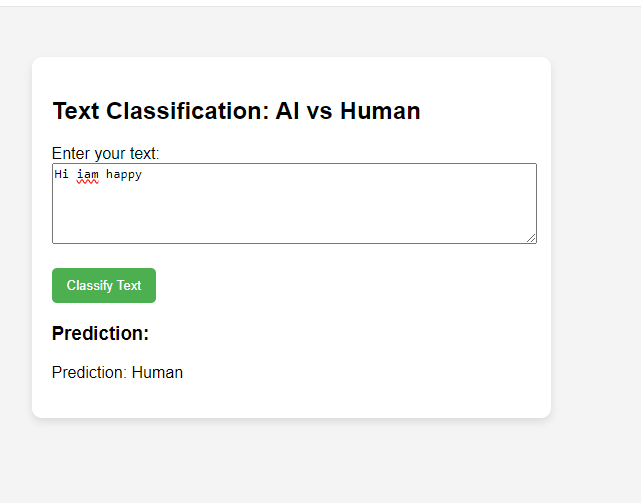
1. **Deployment:**

The models developed in this study were designed to be deployed in real-world environments to classify digital media as either human-generated or AI-generated. The deployment process involved selecting the best-performing model based on evaluation metrics such as accuracy, precision, recall, and F1-score. Once the optimal model was identified, it was integrated into a system that could process and classify new, unseen content in real-time. This system was deployed through an HTML interface, providing users with an accessible platform to interact with the model and receive immediate classification results for any digital media submitted.

1. **AI Content**

****

1. **Human Content**



1. **Conclusion:**

This study successfully developed and evaluated deep learning models (LSTM and CNN) and a traditional machine learning technique (SVM) for detecting AI-generated content. The results indicate that LSTM and CNN models outperform SVM in classifying human- vs. AI-generated text. These findings contribute to the broader field of AI content detection and offer insights into the strengths and limitations of various machine learning models for this critical task. As AI continues to advance, ongoing research and development of more accurate detection systems are essential to preserve trust in digital media and protect against the risks of AI-driven content manipulation. Future work should focus on improving model robustness, addressing class imbalances, and integrating new AI detection technologies to stay ahead of the rapidly evolving capabilities of AI content generation.

**References:**

1. R. Merine and S. Purkayastha, “Risks and benefits of AI-generated text summarization for expert level content in Graduate Health Informatics,” 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI), Jun. 2022. doi: 10.1109/ichi54592.2022.00113.
2. A. M. Elkhatat, K. Elsaid, and S. Almeer, “Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text,” International Journal for Educational Integrity, vol. 19, no. 1, Sep. 2023. doi: 10.1007/s40979-023-00140-5.
3. F. R. Elali and L. N. Rachid, “AI-generated research paper fabrication and plagiarism in the scientific community,” Patterns, vol. 4, no. 3, p. 100706, Mar. 2023. doi: 10.1016/j.patter.2023.100706.
4. G. Gritsay, A. Grabovoy, and Y. Chekhovich, “Automatic detection of machine generated texts: Need more tokens,” 2022 Ivannikov Memorial Workshop (IVMEM), Sep. 2022. doi: 10.1109/ivmem57067.2022.9983964.
5. S. Göring, R. R. Ramachandra Rao, R. Merten, and A. Raake, “Analysis of appeal for realistic AI-generated photos,” IEEE Access, vol. 11, pp. 38999–39012, 2023. doi: 10.1109/access.2023.3267968.
6. C. Li et al., “Agiqa-3K: An open database for AI-Generated Image Quality Assessment,” IEEE Transactions on Circuits and Systems for Video Technology, pp. 1–1, 2024. doi: 10.1109/tcsvt.2023.3319020.
7. H. Du et al., “Exploring collaborative distributed diffusion-based AI-generated content (AIGC) in wireless networks,” IEEE Network, pp. 1–8, 2024. doi: 10.1109/mnet.006.2300223.
8. N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using Siamese Bert-Networks,” Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019. doi: 10.18653/v1/d19-1410.
9. S. Gehrmann, H. Strobelt, and A. Rush, “GLTR: Statistical detection and visualization of generated text,” Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2019. doi: 10.18653/v1/p19-3019.
10. Y. Wang, Y. Pan, M. Yan, Z. Su, and T. H. Luan, “A survey on CHATGPT: AI–generated contents, challenges, and solutions,” IEEE Open Journal of the Computer Society, vol. 4, pp. 280–302, 2023. doi: 10.1109/ojcs.2023.3300321.
11. Y. Lin et al., “Blockchain-aided secure semantic communication for AI-generated content in Metaverse,” IEEE Open Journal of the Computer Society, vol. 4, pp. 72–83, 2023. doi: 10.1109/ojcs.2023.3260732.
12. H. Alamleh, A. A. AlQahtani, and A. ElSaid, “Distinguishing human-written and CHATGPT-generated text using machine learning,” 2023 Systems and Information Engineering Design Symposium (SIEDS), Apr. 2023. doi: 10.1109/sieds58326.2023.10137767.
13. J. P. Wahle, T. Ruas, S. M. Mohammad, N. Meuschke, and B. Gipp, “AI usage cards: Responsibly reporting AI-generated content,” 2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Jun. 2023. doi: 10.1109/jcdl57899.2023.00060.
14. T. T. Nguyen, A. Hatua, and A. H. Sung, “How to detect AI-generated texts?,” 2023 IEEE 14th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), Oct. 2023. doi: 10.1109/uemcon59035.2023.10316132.
15. I. Katib, F. Y. Assiri, H. A. Abdushkour, D. Hamed, and M. Ragab, “Differentiating chat generative pretrained transformer from humans: Detecting chatgpt-generated text and human text using machine learning,” Mathematics, vol. 11, no. 15, p. 3400, Aug. 2023. doi: 10.3390/math11153400.