E-ISSN: 1115 – 8492 P-ISSN: 3027 – 0693
MPRI-JNSAR | INTERNATIONAL JOURNAL OF NATURE AND SCIENCE ADVANCE RES.
VOL. 10 NO. 9 OCTOBER-2025

DETECTION OF AI-GENERATED TEXT THROUGH CLASSICAL MACHINE LEARNING AND DEEP LEARNING APPROACHES

DAMOYE T.
Department of Computer Science, Babcock University, Nigeria.
Corresponding Author:
damoye.temidayo@gmail.com
DOI: https://doi.org/10.70382/mejnsar.v10i9.070

Abstract

The emergence of artificial intelligence (AI) in digital communication has significantly transformed content generation, particularly through its ability to produce text that closely resembles human writing. This advancement has revolutionized content creation across various online platforms, leading to the pervasive presence of AI-generated text that blurs the lines between human and machine authorship. This study examines the effectiveness of various machine learning and deep learning models in identifying AI-generated essays. A dataset comprising 44,852 labeled essays, consisting of 27,361 human-written and 17,491 AI-generated instances, was utilized for the analysis. Classical models such as Naive Bayes, Logistic Regression, Random Forest, and XGBoost were compared to deep learning models, including Feedforward Neural Networks (FNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks. Furthermore, an ensemble model combining RNN and LSTM through soft voting was developed. The performance of these models was evaluated based on accuracy, precision, recall, and F1-score. Notably, the ensemble model achieved the highest accuracy, reaching 95.0%, and outperformed all individual

Keywords: Artificial Intelligence, AI-generated Text, Text Classification, Machine Learning, Deep Learning, Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), Ensemble Learning, Content Authenticity, Natural Language Processing (NLP), Automated Detection Systems

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

models. The results indicated that deep learning models, particularly LSTM and RNN, significantly surpassed classical approaches. The ensemble model demonstrated superior performance across all evaluation metrics, showcasing the advantages of hybrid methodologies. This research highlights the feasibility of automated detection systems and sets a benchmark for future advancements in AI text detection.

Introduction

As companies began to harness artificial intelligence (AI) across various fields, a new paradigm emerged, known as Generative AI. Generative AI focuses on the creation of new and unique content derived from training data (Gupta, Nair, Mishra, Ibrahim, & Bhardwaj, 2024). Generative AI can produce a variety of content, including literature, images, music, and videos. It also encompasses concepts such as film plots and academic papers, allowing for a broad range of applications. One notable application of generative AI is the development of chatbots. Companies such as DeepMind, OpenAI, Google, and Meta are among those that have developed these conversational agents. The popularity of ChatGPT has surged since OpenAI made its user interface publicly accessible. While these technologies are still in their early stages, academic research is beginning to focus on them, and experts are routinely assessing their reliability (Gupta, Nair, Mishra, Ibrahim, & Bhardwaj, 2024).

ChatGPT is currently utilized across various industries and possesses tremendous potential. While some studies have examined its application in education for academic writing and the creation of instructional materials and virtual tours for students, others have investigated its use in healthcare, legal writing, coding, and entertainment (Wang, Scells, Zuccon, & Koopman, 2023); (Zhai, 2023); (Shope, 2023), (Avila-Chauvet, Mejía, & Acosta Quiroz, 2023), (Mijwil & Aljanabi, 2023), (Aydin, Karaarslan, & Damar, 2023).

The advent of artificial intelligence and the emergence of language models, such as ChatGPT, have brought about significant transformations in the marketing sector. ChatGPT is regarded as a transformative tool in the industry, as it empowers marketers to create tailored content that resonates with their intended audience. This technology can be employed to develop virtual assistants for customer service, enabling consumers to receive prompt and personalized responses to their inquiries. Additionally, it facilitates the creation of customized marketing materials, enabling businesses to engage with customers in unique and distinctive ways (Dans, 2023). However, the integration of ChatGPT and other AI technologies is not straightforward, and various theories have emerged to explain why certain companies are more inclined to adopt these innovations (Gupta, Nair, Mishra, Ibrahim, & Bhardwaj, 2024).

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

As the user base of ChatGPT expands, reactions to technology become more pronounced, reflecting broader sentiments about innovation (Dans, 2023). While some claim that ChatGPT is a premier AI chatbot for the general public (Roose, 2022), critics express concerns about the potential negative impacts on its use in the knowledge sector (Krugman, 2022), democratic processes (Cowen, 2022), and ethical considerations associated with its use (Cotton, Cotton, & Shipway, 2023). These dual perspectives highlight the nuanced discourse surrounding ChatGPT's role in industry and society (Gupta, Nair, Mishra, Ibrahim, & Bhardwaj, 2024).

The education sector is undergoing significant changes as LLMs provide opportunities to automate processes such as content creation, grading, and personalized learning.(Wollny et al., 2021), (Mallek et al., 2024), (Analytica, 2023), (Shrivastava, Pupale, & Singh, 2021). Recently, there have been significant advancements in large language models (LLMs), such as ChatGPT (OpenAI, 2022) and ChatGLM (THUDM, 2023). Although these models have demonstrated remarkable general abilities (Bang et al., 2023), (Guo et al., 2023), they still face significant challenges, including factual hallucination (Cao, Dong, Wu, & Cheung, 2020), (Raunak, Menezes, & Junczys-Dowmunt, 2021), (Ji et al., 2023), knowledge obsolescence (He, Zhang, & Roth, 2022), and a lack of domain-specific expertise (Li et al., 2023), (Shen, Chen, Backes, & Zhang, 2023). However, these advancements also raise critical ethical and practical challenges, such as data privacy, bias, overreliance on AI, and pedagogical concerns (Zheng, Zhang, & Woodland, 2021), (Devlin et al., 2018), (Olney, 2023).

Our communication practices are quickly changing due to the emergence of generative AI models. It is widely used in various disciplines, including healthcare, academic research, the arts, and content production (Najjar, Ashqar, & Darwish, 2025). Large Language Models (LLMs) have demonstrated impressive capabilities in understanding user inquiries and generating text that closely resembles human speech. As a result, LLMs have garnered considerable attention from researchers, policymakers, and educators alike. While these models are often heralded for their potential to transform society, they also pose certain risks. The emergence of innovative AI-based chatbots highlights the importance of questioning the originality of ideas, language, and solutions, specifically whether a sentence originated from an AI or human source. Analyzing the originality of written content has significant implications across various fields, including digital forensics and information security (Cao et al., 2023). Large Language Models (LLMs) have demonstrated impressive capabilities in understanding user inquiries and generating text that closely resembles human speech. As a result, LLMs have garnered considerable attention from researchers, policymakers, and educators alike. While these models are

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

often heralded for their potential to transform society, they also pose certain risks. The emergence of innovative AI-based chatbots highlights the importance of questioning the originality of ideas, language, and solutions, specifically whether a sentence originated from an AI or human source. Analyzing the originality of written content has significant implications across various fields, including digital forensics and information security.(Qadir, 2023). Furthermore, large language models (LLMs) can produce inaccurate responses and information due to their reliance on outdated data or the potential for generating hallucinations (Cao et al., 2023), (Ashqar et al., 2024). As generative AI tools like ChatGPT become increasingly prevalent in educational settings, educators are increasingly concerned about ensuring that student work is genuinely authored by humans (Tami, Ashqar, & Elhenawy, 2024), (Masri et al., 2024), (Rouzegar & Makrehchi, 2024). This study aims to develop a robust model for detecting AI-generated text, providing educators with a valuable resource that promotes fairness and upholds academic integrity.

The broader educational significance of this research lies in its implications for plagiarism detection, the assessment of digital submissions, and the protection of learning outcomes in environments where AI-generated content is increasingly prevalent. By contributing to the understanding of how digital technologies can enhance both pedagogy and ethical standards in education, this study addresses the issue of plagiarism among students who use these tools to complete homework and term papers. Many academic programs emphasize the importance of strong writing skills.

History of Generative AI

AI Generative models have a long history in artificial intelligence, dating back to the 1950s with the development of Hidden Markov Models (HMMs) (Knill & Young, 1997) and Gaussian Mixture Models (GMMs) (Reynolds et al., 2009). These models generated sequential data such as speech and time series. However, it wasn't until the advent of deep learning that generative models saw significant improvements in performance. In the early years of deep generative models, different areas generally had little overlap. In natural language processing (NLP), a traditional method for generating sentences involves learning word distributions using N-gram language modeling (Bengio, Ducharme, Vincent, & Janvin, 2003) and then searching for the best sequence. However, this method cannot effectively adapt to long sentences. To solve this problem, recurrent neural networks (RNNs) (Mikolov et al., 2010) were later introduced for language modeling tasks, allowing for modeling relatively long dependencies. This was followed by the development of Long Short-Term Memory (LSTM) (Graves, 2012) and Gated Recurrent

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

Unit (GRU) (Dey & Salem, 2017), which leveraged a gating mechanism to control memory during training. These models are capable of attending to around 200 tokens in a sample (Khandelwal, He, Qi, & Jurafsky, 2018), which marks a significant improvement compared to N-gram language models. In the realm of computer vision (CV), prior to the introduction of deep learning-based techniques, traditional image generation methods relied on algorithms such as texture synthesis (Efros & Leung, 1999) and texture mapping (Heckbert, 1986). These algorithms were grounded in hand-crafted features and faced limitations in their capacity to produce complex and varied images. The proposal of Generative Adversarial Networks (GANs) in 2014 (Goodfellow et al., 2014) marked a significant milestone in this field, yielding impressive outcomes across a range of applications. Additionally, Variational Autoencoders (VAEs) (Kingma & Welling, 2013) and diffusion generative models (Song & Ermon, 2019) have been developed to provide finer control over the image generation process, resulting in the production of high-quality images. The advancement of generative models has taken varied paths across different areas; however, a notable convergence occurred with the emergence of the transformer architecture (Vaswani et al., 2017). Introduced by Vaswani et al. for natural language processing (NLP) tasks in 2017, transformers have since been applied to computer vision, eventually establishing themselves as the dominant backbone for numerous generative models across various domains (Ramesh et al., 2021), (Lewis et al., 2019). In NLP, many influential large language models, such as BERT and GPT, utilize the transformer architecture as their foundational element, delivering advantages over preceding models like LSTM and GRU. In the CV domain, the Vision Transformer (ViT) (Dosovitskiy et al., 2020) and Swin Transformer (Liu et al., 2021) have further advanced this concept by integrating the
transformer architecture with visual components, enabling their application to image-based tasks.

Except for the improvement that transformer brought to individual modalities, this intersection also enabled models from different domains to be fused together for multimodal tasks. One such example of multimodal models is CLIP (Radford et al., 2021). CLIP is a joint vision-language model that combines the transformer architecture with visual components, allowing it to be trained on a massive amount of text and image data. Since it combines visual and language knowledge during pre-training, it can also be utilized as an image encoder in multimodal prompting for generation. In all, the emergence of transformer-based models has revolutionized AI generation, leading to the possibility of large-scale training. In recent years, researchers have also begun to introduce new techniques based on these models. For instance, in NLP, instead of fine-

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

tuning, people sometimes prefer few-shot prompting (Dong et al., 2023), which involves including a few examples selected from the dataset in the prompt to help the model better understand the task requirements. In visual language, researchers often combine modality-specific models with self-supervised contrastive learning objectives to provide more robust representations.

Classical Machine Learning Approaches

Classical machine learning models have been widely applied in text classification tasks. Models such as Multinomial Naive Bayes, Logistic Regression, Random Forest, and XGBoost operate on engineered features derived from vectorization techniques like Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) (Oancea, 2023). These models are computationally efficient and easy to interpret. However, they struggle with contextual understanding and semantic nuances, limiting their effectiveness against advanced AI-generated texts (Watts & Adriano, 2020).

Deep Learning Approaches

Deep learning models significantly enhance the detection of complex patterns in text. These models excel over traditional methods in numerous text classification tasks due to their capacity to process contextual information. However, they require large datasets and substantial computational resources (Minaee et al., 2021)

Machine-Generated Text Detection Task

The machine-generated text detection task consists of automatically detecting content generated by LLMs. It can be solved as a binary classification problem or by setting a threshold. From a mathematical perspective, it can be formalized as a binary classification problem, seeking to determine if a given text is generated by an LLM or by a human writer (Watts & Adriano, 2020)

Given a text t and a Detector $D(t)$, the equation is the following:

$$D(t) = \begin{cases} 1 & \text{if } t \text{ is machine} - \text{generated,} \\ 0 & \text{if } t \text{ is human} - \text{written} \end{cases}$$

Setting a threshold can contribute to another way to define the detection task. Given an input text, the text detector outputs a score. A score higher than the threshold indicates a machine-generated text (Ghosal et al., 2023)

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

Large Language Models (LLMs) are built on advanced neural networks, particularly utilizing the Transformer architecture introduced by Vaswani et al. It is based on the following equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Each word in the sentence is converted into a numerical vector. Q (query) is the word vector, K (key) are the vectors for every word in the sentence, and V (value) are the value vectors. The softmax function converts scores into a probability distribution. This formula allows the model to weigh the importance of each word within a sequence of other words. Generation of text (contextually relevant and coherent) by LLMs like Xlnet depends on their autoregressive nature. Autoregressive models predict future behavior based on past behavior data. In the case of text, the subsequent word is predicted based on the previous ones (Yang, 2019). The field of generated text detection has become a prominent area of research. Numerous studies have reviewed the primary techniques employed in this domain, with one of the earliest reviews dating back to 2016 (Beresneva, 2016). Since then, the rapid proliferation of large language models (LLMs) has significantly complicated the detection of machine-generated text. As a result, researchers have increasingly focused on analyzing text produced by these models, leading to the development and publication of new detection methods (Jawahar, Abdul-Mageed, & Lakshmanan, 2020). Following the rise of ChatGPT, there has been an even greater surge in reviews of LLM-generated text, including notable approaches proposed by Dhaini et al (Dhaini, Poelman, & Erdogan, 2023). This research direction has emerged as a central theme, bolstered by numerous studies that have also influenced the current work. In detail, a further step has been made by a work that introduces a first kind of categorization by dividing feature-based and neural-language model approaches (Crothers, Japkowicz, & Viktor, 2023); another work (Tang, Chuang, & Hu, 2024) divides the task into black-box and white-box detection, introducing a novel template followed by successive works; another method (Yang et al., 2023), following the black-box and white-box structure idea, added three categories of detection methods: training-based, zero-shot, and watermarking methods. Moreover, interesting research (Ghosal et al., 2023) highlighted the weaknesses of existing text detection techniques (e.g, text paraphrasing). At the same time, the review work by Uchendu et al. (Uchendu, Le, & Lee, 2023) has introduced the hybrid methods category for the first time. The most exhaustive study, considering the state-of-the-art methods to date, is the one proposed by Wu et al

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

(Wu et al., 2023), which covers many method categories. It presents applicable sections for those dealing with the phenomenon of generated text detection, providing details regarding the most popular datasets and benchmarks useful for this task, and highlighting the research limitations in generated text detectors.

Exploratory Data Analysis

Distribution of Essay Length by Source (Human vs. AI)

I analyzed patterns in essay length and word count to explore how human authors and AI systems compose text. A histogram accompanied by Kernel Density Estimation (KDE) illustrates the distribution of each group across various length categories.

The resulting plot reveals a significant difference: AI-generated essays tend to cluster closely around a specific word count, indicating that these models consistently produce text of similar lengths. This level of consistency suggests that the engines have a clear understanding of how long a finalized piece should be.

In contrast, human-written essays display a broader range of lengths, reflecting the varied rhythms, moods, and skill levels that real writers exhibit. Due to this difference, machine learning systems can leverage length data as a reliable indicator when determining whether a passage was authored by a human or an AI.

This difference in distribution is a helpful feature for machine learning models attempting to distinguish between the two classes.

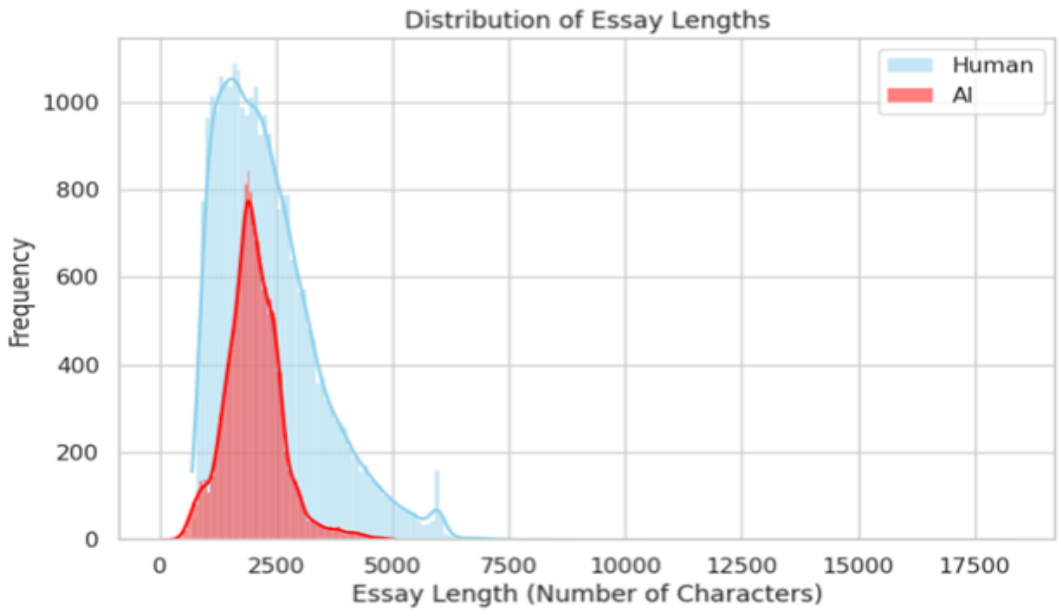


Figure 1: Distribution of Essay Lengths

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

Comparative Analysis of Essay Length Using Boxplots

To further explore differences in writing behaviour, I used boxplots to visualise the distribution of essay lengths for both human and AI-generated texts.

Key observations from the plot:

- Median Length: The median essay length for human-written texts is slightly lower than that of AI-generated essays. This aligns with the observation that AI tends to produce more uniformly lengthy responses.
- Interquartile Range (IQR): AI-generated essays exhibit a narrower IQR, suggesting that most AI essays fall within a tight range of lengths. On the other hand, human-written essays display a wider IQR, showing greater diversity in response lengths.
- Outliers: Human essays include significantly more outliers, reflecting the tendency of humans to write either very short or very long essays, likely depending on their interpretation of the prompt or writing style or if they are bounded by word counts as some essays would specify. AI-generated texts, being more regulated, rarely exhibit such extremes.

This analysis further emphasises the structural predictability of AI writing compared to the diverse and variable nature of human expression.

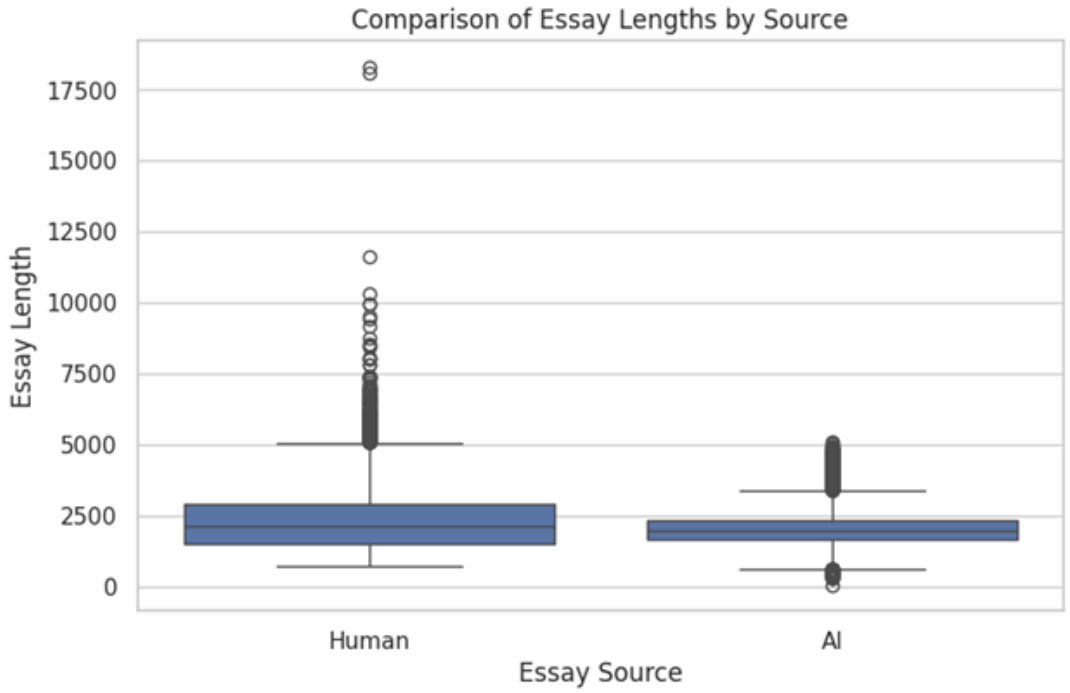


Figure 2: Comparison of Essay Length

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

Dataset Description

The dataset used in this study comprises a total of 44,868 persuasive-style essays, including both **human-written** and **AI-generated** texts. It was originally uploaded to Kaggle by researcher Darek Kleczek and curated from multiple reputable sources.

- **Human-written essays (~26,000):** These were sourced from the publicly available Persuasive Essay Corpus, which includes argumentative essays written by U.S. students in grades 6–12 in response to 15 writing prompts (e.g., Phones and driving).
- **AI-generated texts (~17,000):** These were produced by 17 different large language models (LLMs), including both proprietary and open-source models such as GPT-3.5-turbo, GPT-4, LLaMA2, Falcon 180B, and MPT. These samples were scraped or downloaded from curated platforms like Kaggle and other open-access repositories.

Each record in the dataset retains the following key variables:

- **Source:** Indicates the origin of the text (e.g., persuade_corpus or the specific LLM used).
- **Label:** Binary classification label identifying the authorship of the text:
  - 0 = Human-written
  - 1 = AI-generated
- **Text:** The full essay body or written content.
- **Prompt_name:** The specific writing prompt or topic that guided the essay (e.g., "Phones and driving").

To ensure robust evaluation, the dataset was randomly split into:

- Training set (80%) – Used to train the classification models.
- Test set (20%) – Held out for performance evaluation and generalization testing.

The inclusion of texts from diverse language models enhances the dataset's realism, mimicking real-world scenarios where AI-generated content varies in style and quality.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

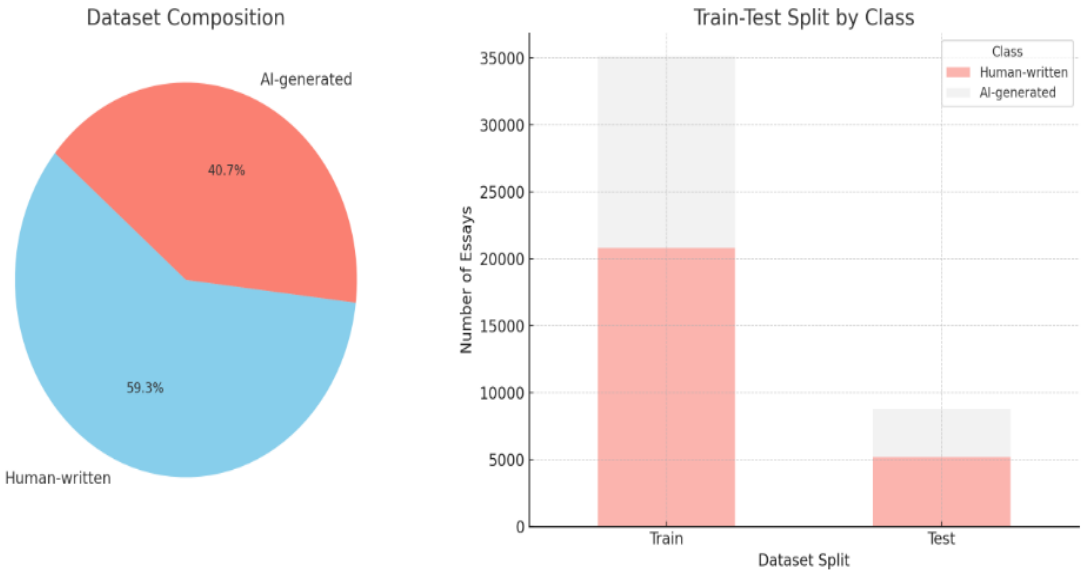


Figure 3: Dataset Composition Pie Chart and Train-Test Split Bar Chart

Data Cleaning

To prepare the data for analysis, a normalization function was applied to the text column to standardise formatting: punctuation was removed, newline characters were replaced with whitespace, and excess whitespace was trimmed. No missing values were present in the dataset, confirming completeness and readiness for modeling

Data Preprocessing

Effective preprocessing is a vital phase in transforming raw text data into a format suitable for machine learning models. This step ensures that the data is clean, consistent, and ready for both training and evaluation. The preprocessing pipeline for this study included the following key components:

Dataset Splitting

The dataset was split using the standard **80-20 ratio**, where:

- **80%** was allocated for training.
- **20%** was held out for testing to evaluate model performance on unseen data.

Importantly, the stratify parameter was used during splitting to maintain the original class distribution (i.e., human-written vs. AI-generated texts) across both subsets. This was

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

crucial due to class imbalance and ensured both classes were adequately represented. Splitting was performed prior to any feature extraction or scaling to avoid data leakage, which could otherwise result in overly optimistic performance metrics.

Text Cleaning and Normalization

To reduce noise and standardise inputs, the following text cleaning steps were applied:

- Removal of HTML tags, special characters, digits, and excess whitespace
- Lowercasing all text to ensure uniformity
- Stopword removal using NLTK's English stopword list to eliminate uninformative common words
- Tokenization and Lemmatization to break text into tokens and reduce them to their base forms, helping preserve semantic meaning and reduce sparsity

Text Vectorization

Since machine learning models require numerical input, the cleaned text was converted into numerical feature representations using different approaches based on the model type:

For Classical Machine Learning Models:

TF-IDF (Term Frequency-Inverse Document Frequency) vectorization was applied with the following parameters:

- ngram_range=(3, 4): Captures trigrams and quadrigrams to better represent phrases and local context.
- tokenizer=lambda x: re.findall(r'[^\W]+', x): A custom tokenizer using regular expressions to extract word-like tokens, ignoring punctuation.
- stop_words='english': Removes common English words unlikely to contribute to meaningful classification.
- strip_accents='unicode': Standardised text by removing accent marks (e.g., "crème" becomes "creme").
- sublinear_tf=True: Applies logarithmic scaling to term frequency to mitigate the impact of very frequent terms.
- max_features=50,000: Limits the feature space to the top 50,000 most frequent n-grams, balancing performance and computational efficiency.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

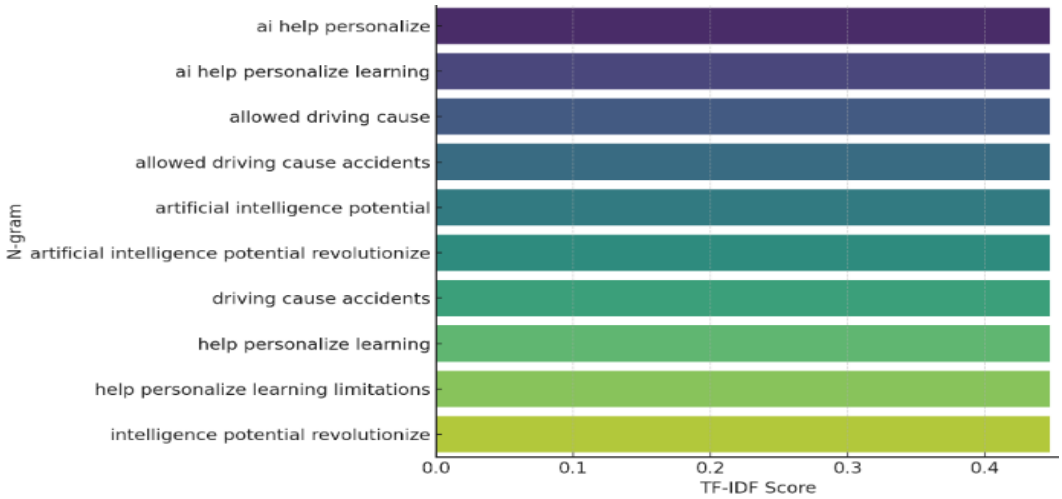


Figure 4: **Top 10 Trigrams and Quadrigrams based on their TF-IDF scores**

For Deep Learning Models:

- GloVe embeddings were used to represent words in dense vector space, capturing semantic relationships between words.
- Padding and Truncation: Input sequences were padded or truncated to a fixed length (e.g., 300 tokens) to ensure uniform input size required by neural models.

This structured preprocessing pipeline enabled effective model training, reduced noise, and ensured that the classification algorithms had high-quality, meaningful inputs for distinguishing between human-written and AI-generated essays.

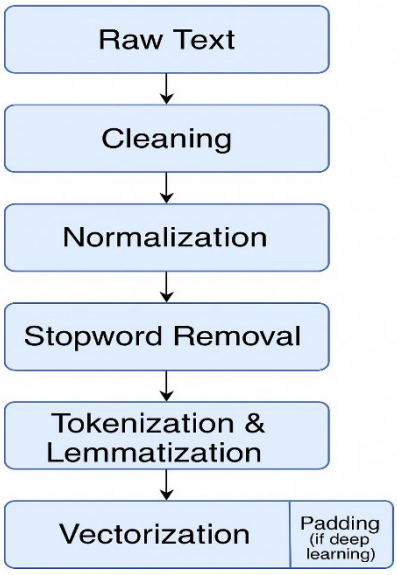


Figure 5: Flowchart of preprocessing pipeline

Experimental Setup

The dataset consisted of 44,852 labeled essays (27,361 human-written and 17,491 AI-generated), split into 80% training and 20% testing sets. Preprocessing steps included tokenization, stopword removal, and lemmatization. Classical models were trained on TF-IDF and BoW features, while deep learning models utilised word embeddings. All experiments were conducted in Python using TensorFlow, Keras, and scikit-learn libraries.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

Performance of Classical Machine Learning Models and Deep Learning Models

The performance and ROC-AUC scores of the four classical machine learning models and deep learning models are presented in Table 4.1 and Table 4.2, respectively. Among classical models, XGBoost emerged as the top performer. Its gradient boosting mechanism allowed it to learn complex patterns in feature space, outperforming other models in all metrics. Random Forest also performed well, leveraging ensemble decision trees. Multinomial Naive Bayes had the lowest performance, likely due to its assumption of feature independence

LSTM outperformed all individual models due to its ability to capture long-term dependencies and contextual relationships in text. RNN also demonstrated strong performance, effectively extracting local patterns in word sequences. Simple RNN lagged behind, showing limitations in handling vanishing gradients and longer contexts.

The ensemble approach yielded the highest accuracy, precision, recall, and F1-score. This suggests that leveraging complementary strengths of RNN and LSTM enhanced the robustness of classification. RNN captured local n-gram patterns while LSTM provided memory of longer sequences, creating a balanced and powerful hybrid.

Table 1: Performance Metrics of Classical and Deep Learning Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Remarks
Multinomial Naive Bayes	83.75	81.52	86.72	84.04	Performs well on simple features; fast and efficient
Logistic Regression	87.50	86.18	89.26	87.69	Balances performance and interpretability
Random Forest	88.50	86.81	91.08	88.89	Handles high-dimensional data; robust against overfitting
XGBoost	90.50	89.13	91.62	90.36	Best classical model; strong performance across all metrics
Feedforward Neural Network (FNN)	88.75	87.67	89.27	88.46	Learns non-linear relationships; requires tuning
Simple RNN	89.75	88.08	91.54	89.78	Handles sequences but struggles with long dependencies
Long Short-Term Memory (LSTM)	92.25	90.26	93.71	91.95	Excellent for sequential data
Ensemble (RNN + LSTM)	95.00	93.81	96.42	95.10	Highest accuracy; combines spatial and temporal features

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

Table 2: ROC-AUC Scores by Model

Model	ROC-AUC Score	Remarks
Multinomial Naive Bayes	0.83	Fairly good for a simple model
Logistic Regression	0.88	Shows strong discriminatory power
Random Forest	0.90	Consistent with accuracy score
XGBoost	0.91	Reflects robustness of predictions
Feedforward NN	0.90	Learns useful global patterns
Simple RNN	0.91	Good balance between accuracy and recall
LSTM	0.93	Strong performance on sequential detection
Ensemble (RNN + LSTM)	0.96	Best at distinguishing AI vs. human text

Comparison of Model Performance

A visual comparison of all models' performance indicates a clear advantage of deep learning models over classical ones. XGBoost was the best among classical models, closely rivaling the FNN. LSTM and RNN significantly outperformed all others, with the ensemble achieving the best overall results.

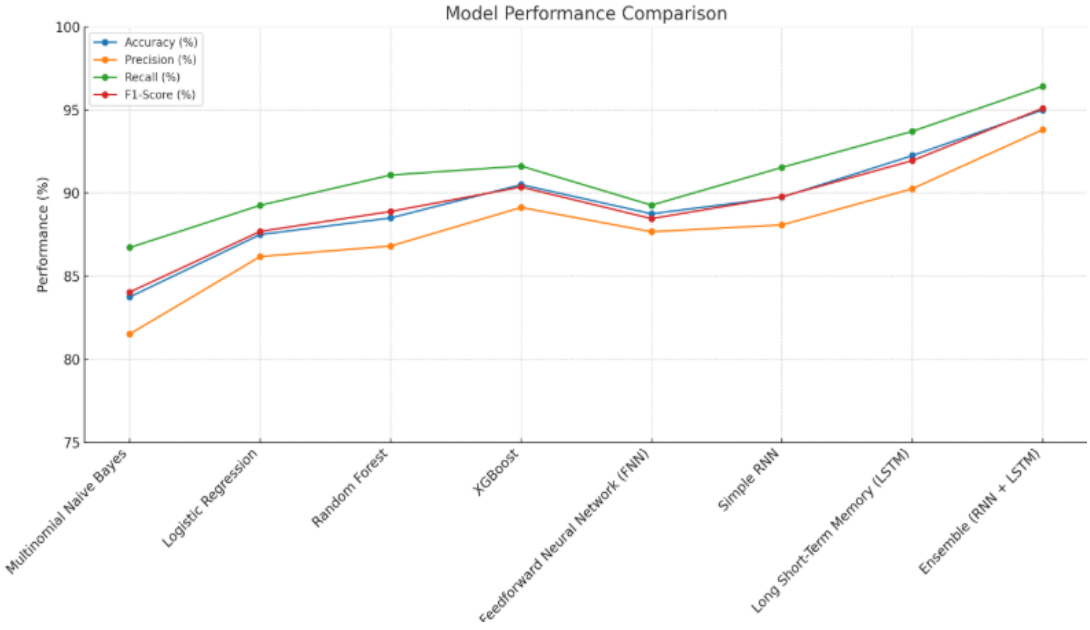


Figure 6: Model Performance Metrics Comparison

Error Analysis Misclassifications

These were analysed to understand potential model weaknesses. Common patterns observed include:

1. Human essays with rigid structure and formal tone often mimicked AI outputs.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

2. AI-generated essays that included personal anecdotes or rhetorical questions misled the models into classifying them as human-written.

Summary

This research investigated the application of classical machine learning and deep learning models for identifying AI-generated content in written texts. With the rise of advanced Large Language Models (LLMs) like ChatGPT, LLaMA, and Falcon, distinguishing between human and AI-written text has become increasingly challenging. The research utilised a comprehensive dataset containing 44,868 essays—comprising both human-written persuasive texts and AI-generated content from various open-source LLMs.

Seven individual classifiers—Multinomial Naive Bayes, Logistic Regression, Random Forest, XGBoost, Feedforward Neural Network (FNN), Simple RNN, and LSTM—were compared, alongside an ensemble model combining RNN and LSTM. The preprocessing pipeline involved text cleaning, normalization, tokenization, and TF-IDF vectorization. Exploratory Data Analysis (EDA) revealed structural differences in essay length and variability between human and AI-generated texts, further enriching model inputs.

Model performance was evaluated using metrics such as accuracy, precision, recall, and F1-score. Among all tested models, the ensemble deep learning model (RNN + LSTM) demonstrated superior performance, achieving an accuracy of 95.00%, F1-score of 95.10%, and a recall of 96.42%. These results confirm the effectiveness of deep learning models, particularly when combined, in handling the nuanced patterns of AI-generated texts.

Conclusion

The findings of this study demonstrate that while classical machine learning models offer commendable results in distinguishing between AI-generated and human-written content, deep learning models especially the ensemble of RNN and LSTM provide a more robust and accurate approach. The research also highlights key differences in writing behaviour between humans and machines, particularly in essay length distribution and consistency, which can be leveraged for classification.

Given the accelerating use of AI-generated content across various domains such as education, journalism, and digital marketing, the need for reliable detection systems is critical. This research contributes to this emerging field by presenting a practical, scalable framework that combines classical ML and deep learning techniques for enhanced content authenticity verification.

Recommendation

Based on the findings of this study, it is recommended that institutions and academic bodies adopt more advanced deep learning models, particularly ensemble approaches

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

combining RNN and LSTM, for the detection of AI-generated content. Educators should be equipped with tools that can identify subtle differences in writing patterns between human and AI-generated texts. Furthermore, ongoing training and awareness campaigns can help students understand the ethical implications of using AI tools in academic work. Efforts should also be made to improve dataset diversity in future research, to ensure model generalisation across writing styles and domains.

Suggestions for Further Studies

1. **Incorporation of Pre-trained Language Models**: Future research should investigate the use of state-of-the-art transformers like BERT, RoBERTa, T5, and ELECTRA for content classification. These models are known for their high performance in NLP tasks and may further improve detection accuracy.
2. **Real-Time Detection Systems**: Developing real-time tools for detecting AI-generated text in live environments, such as classrooms, newsrooms, or social media platforms, is an important next step.
3. **Cross-lingual Studies**: This study focused solely on English essays. Expanding the scope to include multilingual datasets could help build more inclusive and globally relevant detection systems.
4. **Behavioural and Psychological Impact**: Further studies should examine how the presence of AI-generated content affects reader trust, decision-making, and perception particularly in sensitive contexts like news reporting and academia.

Reference

A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in Proceedings of the seventh IEEE international conference on computer vision, vol. 2, pp. 1033–1038, IEEE, 1999.

A. A. Najjar, H. I. Ashqar, and O. A. Darwish, "Detecting AI-Generated Text in Educational Content: Leveraging Machine Learning and Explainable AI for Academic Integrity

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," 2021.

A. Uchendu, T. Le, D. Lee, "Attribution and obfuscation of neural text authorship: A data mining perspective," ACM SIGKDD Explorations Newsletter, vol. 25, no. 1, pp. 1–18, 2023.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, "Attention is all you need," in Proceedings of the 31st International Conference on Neural Information Processing Systems NIPS'17, Red Hook, NY, USA, 2017, p. 6000–6010, Curran Associates Inc.

Avila-Chauvet, L. and Mejía, D., Acosta Quiroz, C.O. (2023) Chatgpt as a support tool for online behavioral task programming. Available at SSRN: Https://ssrn.com/abstract=4329020.

Aydin, Ö., Karaarslan, E., Damar, M. (2023). Is ChatGPT leading generative AI? What is beyond expectations?. Available at: Https://ssrn.com/abstract=4347800.

B. Oancea, "Text classification using machine learning methods," University of Bucharest, Bucharest, Romania. 2023

Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; Do, Q. V.; Xu, Y.; and Fung, P. 2023.A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity.arXiv:2302.04023

Cao, M.; Dong, Y.; Wu, J.; and Cheung, J. C. K. 2020.Factual Error Correction for Abstractive Summarization Models.In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 6251–6258. Online: Association for Computational Linguistics.

Cotton, D.R., Cotton, P.A., & Shipway, J.R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Preprint. 10.35542/osf.io/mrz8h.

D. A. Reynolds et al., "Gaussian mixture models.," Encyclopedia of biometrics, vol. 741, no. 659-663, 2009.

D. Beresneva, "Computer-generated text detection using machine learning: A systematic review," in Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, Proceedings 21, 2016, pp. 421–426, Springer

D. P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114, 2013.Detection.arXiv:2301.07597.

Devlin J, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

E. Crothers, N. Japkowicz, H. Viktor, "Machine-generated text: A comprehensive survey of threat models and detection methods," IEEE Access, vol. PP, pp. 1–1, 01 2023, doi:10.1109/ACCESS.2023.3294090.

E. Dans ChatGPT… y el miedo a la innovación Enrique Dans (2023) http://bit.ly/3ZGGMTb Education: A Comparative Study of GPT-3.5 and GPT-4 in Crafting Customized Test Questions," arXiv preprint arXiv:2406.13903, 2024.

G. Jawahar, M. Abdul-Mageed, V. Laks Lakshmanan, "Automatic detection of machine generated text: A critical survey," in Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 2296–2309.

Guo, B.; Zhang, X.; Wang, Z.; Jiang, M.; Nie, J.; Ding, Y.; Yue, J.; and Wu, Y. 2023.How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection.arXiv:2301.07597.

H. I. Ashqar, A. Jaber, T. I. Alhadidi, and M. Elhenawy, "Advancing Object Detection in Transportation with Multimodal Large Language Models (MLLMs): A Comprehensive Review and Empirical Testing," arXiv preprint arXiv:2409.18286, 2024.

H. Rouzegar and M. Makrehchi, "Generative AI for Enhancing Active Learning in Education: A Comparative Study of GPT-3.5 and GPT-4 in Crafting Customized Test Questions," arXiv preprint arXiv:2406.13903, 2024.

He, H.; Zhang, H.; and Roth, D. 2022.Rethinking with Retrieval: Faithful Large Language Model Inference.arXiv:2301.00303.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

J. Qadir, "Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education," in 2023 IEEE Global Engineering Education Conference (EDUCON), IEEE, 2023, pp. 1–9.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

J. Watts and A. Adriano, "Uncovering the Sources of Machine-Learning Mistakes in Advertising: Contextual Bias in the Evaluation of Semantic Relatedness," J. Advert., vol.50, no. 1, pp. 1–13, Oct. 2020, doi: 10.1080/00913367.2020.1821411

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023.Survey of Hallucination in Natural Language Generation.ACM Comput. Surv., 55(12).

K. Knill and S. Young, "Hidden Markov Models in Speech and Language Processing," in Corpus-Based Methods in Language and Speech Processing (S. Young and G. Bloothooft, eds.), pp. 27–68, Dordrecht: Springer Netherlands, 1997.

K. Roose The brilliance and weirdness of ChatGPT The New York Times (2022) http://bit.ly/3XcGlhI

K. S. Kalyan, "A survey of GPT-3 family large language models including ChatGPT andGPT-4," Natural Language Processing Journal, p. 100048, 2023.

Li, X.; Zhu, X.; Ma, Z.; Liu, X.; and Shah, S. 2023c.Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? An Examination on Several Typical Tasks.arXiv:2305.05862.

M. Abu Tami, H. I. Ashqar, M. Elhenawy, S. Glaser, and A. Rakotonirainy, "Using Multimodal Large Language Models (MLLMs) for Automated Detection of Traffic Safety-Critical Events," Vehicles, vol. 6, no. 3, pp. 1571–1590, 2024

M. Dhaini, W. Poelman, E. Erdogan, "Detecting chatgpt: A survey ofthe state of detecting chatgpt-generated text," in Proceedings of the 8th Student Research Workshop associated with the International Conference Recent Advances in Natural Language Processing, 2023, pp. 1–12.

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," arXiv preprint arXiv:1910.13461, 2019.

M. Mijwil, M. Aljanabi Towards artificial intelligence-based cybersecurity: The practices and ChatGPT generated ways to combat cybercrime Iraqi Journal For Computer Science and Mathematics, 4 (1) (2023), pp. 65-70

M. Tami, H. I. Ashqar, and M. Elhenawy, "Automated Question Generation for Science Tests in Arabic Language Using NLP Techniques," arXiv preprint arXiv:2406.08520, 2024.

Mallek F, et al. A review on cultivating effective learning: synthesizing educational theories and virtual reality for enhanced educational experiences. PeerJ Comput Sci. 2024;10: e2000. https://doi.org/10.7717/peerj-cs.2000. Analytica O. GPT-4 underlines mismatch on AI policy and innovation. Emerald Expert Briefings, 2023(oxan-es).

Olney AM. Generating multiple choice questions from a textbook: LLMs match human performance on most metrics. in AIED Workshops. 2023.

OpenAI. 2022.Chatgpt: Optimizing language models for dialogue.https://openai.com/blog/chatgpt.

P. Krugman Does chatgpt mean robots are coming for the skilled jobs? December 6th The New York Times (2022) http://bit.ly/3HdnAp2

P. S. Heckbert, "Survey of texture mapping," IEEE computer graphics and applications, vol. 6, no. 11, pp. 56–67, 1986.

Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, L. Li, and Z. Sui, "A survey on in-context learning," 2023

R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (gru) neural networks," in 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS), pp. 1597–1600, IEEE, 2017.

R. Gupta, K. Nair, M. Mishra, B. Ibrahim, and S. Bhardwaj, "Adoption and impacts of generative artificial intelligence: Theoretical underpinnings and research agenda," International Journal of Information Management Data Insights, vol. 4, no. 1, p. 100232, Apr. 2024, doi: 10.1016/j.jjimei.2024.100232.

R. Tang, Y.-N. Chuang, X. Hu, "The science of detecting llm- generated text," Communications of the ACM, vol. 67, p. 50–59, mar 2024, doi:10.1145/3624725.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

Raunak, V.; Menezes, A.; and Junczys-Dowmunt, M. 2021.The Curious Case of Hallucinations in Neural Machine Translation.In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1172–1183. Online: Association for Computational Linguistics.

S. Ghosal, S. Chakraborty, J. Geiping, F. Huang, D. Manocha, A. Bedi, "A survey on the possibilities & impossibilities of ai-generated text detection," Transactions on Machine Learning Research, 2023

S. Jaradat, R. Nayak, A. Paz, H. I. Ashqar, and M. Elhenawy, "Multitask Learning for Crash Analysis: A FineTuned LLM Framework Using Twitter Data," Smart Cities, vol. 7, no. 5, pp. 2422–2465, 2024, doi: 10.3390/smartcities7050095.

S. Masri, Y. Raddad, F. Khandaqji, H. I. Ashqar, and M. Elhenawy, "Transformer Models in Education: Summarizing Science Textbooks with AraBART, MT5, AraT5, and mBART," arXiv preprint arXiv:2406.07692, 2024.

S. Minaee et al., "Deep Learning--based Text Classification: A Comprehensive Review," ACM Comput. Surv., vol. 54, no. 3, pp. 1–40, Apr. 2021, doi: 10.1145/3439726.

Shen, X.; Chen, Z.; Backes, M.; and Zhang, Y. 2023.In ChatGPT We Trust? Measuring and Characterizing the Reliability of ChatGPT.arXiv:2304.08979

Shope, M. (2023). Best practices for disclosure and citation when using artificial intelligence tools. Available at SSRN: Https://ssrn.com/abstract=4338115.

Shrivastava A, Pupale R, Singh P. Enhancing aggression detection using GPT-2 based data balancing technique. in 2021 5th International Conference on Intelligent Computing and control systems (ICICCS). 2021. IEEE.

T. Cowen ChatGPT could make democracy even more messy Bloomberg (2022) http://bit.ly/3XwqjPs

T. Mikolov, M. Karafiát, L. Burget, J. Cernocky, and S. Khudanpur, "Recurrent neural network based language model.," ` in Interspeech, vol. 2, pp. 1045–1048, Makuhari, 2010. A. Graves and A. Graves, "Long short-term memory," Supervised sequence labelling with recurrent neural networks, pp. 37–45, 2012.

THUDM. 2023b.ChatGLM2-6B.https://github.com/THUDM/ChatGLM2-6B.

U. Khandelwal, H. He, P. Qi, and D. Jurafsky, "Sharp nearby, fuzzy far away: How neural language models use context," 2018.

Wang, S., Scells, H., Zuccon, G., & Koopman, B. (2023). Can ChatGPT write a good boolean query for systematic review literature search?. arXIV preprint arXIV:2302.03495v3.

Wollny S, et al. Are we there yet? A systematic literature review on chatbots in education. Front Artif Intell. 2021;4: 654924.

Wu, S. Yang, R. Zhan, Y. Yuan, D. F. Wong, L. S. Chao, "A survey on llm-generated text detection: Necessity, methods, and future directions," ArXiv, vol. abs/2310.14724, 2023

X. Yang, L. Pan, X. Zhao, H. Chen, L. R. Petzold, W. Y. Wang, W. Cheng, "A survey on detection of llms-generated content," ArXiv, vol. abs/2310.15654, 2023

Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," J. Mach. Learn. Res., vol. 3, p. 1137–1155, mar 2003.

Y. Cao et al., "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt," arXiv preprint arXiv:2303.04226, 2023.

Y. Shen, L. Heacock, J. Elias, K.D. Hentel, B. Reig, G. Shih, et al. ChatGPT and other large language models are double-edged swords Radiology (2023), Article 230163

Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," Advances in Neural Information Processing Systems, vol. 32, 2019.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF international conference on computer vision, pp. 10012–10022, 2021.

Z. Yang, "Xlnet: Generalized autoregressive pretraining for language understanding," arXiv preprint arXiv:1906.08237, 2019

Zhai, X. (2023). ChatGPT for next generation science learning. Available at SSRN 4331313.

Zheng X, Zhang C, Woodland PC. Adapting GPT, GPT-2 and BERT lan- guage models for speech recognition. in 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). 2021. IEEE.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]