Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 8 – No. 1, January 2026 – www.caeaccess.org

Analyzing Word Frequency and Predictive Patterns in AI-Generated Essays

Henry Sanmi Makinde
University of North Carolina Greensboro, US
Educational Research Methodology Department

Akindeji Ibrahim Makinde
The Federal University of Technology, Akure
Information Systems Department

Mutiyat Adeola Usman
North Carolina A&T State University
Data Science and Analytics Department

Hope Adegoke
University of North Carolina Greensboro, US
Educational Research Methodology Department

ABSTRACT
Artificial Intelligence (AI) has dramatically transformed various aspects of human life and activities, including the composition of essays and texts. AI technologies have enabled computers to generate text that closely resembles human writing and this has raised concerns with implications for academic integrity, creative authenticity, and professional communication. This study aim to investigates the linguistic characteristics and predictive mechanisms underlying AI-generated essays, aiming to identify markers that distinguish them from human-authored texts. 1,000 essays with diverse topics and writing styles were generated using ChatGPT, DeepSeek, and Gemini and a comparable corpus of human-written essays were also collected from publicly available sources. The research work used natural language processing (NLP) techniques and machine learning models to analyze word frequency, next-word prediction patterns, and stylistic elements in a corpus of AI-generated and human-written essays.The results show that the temperature settings in AI models significantly influence word selection, with higher temperatures increasing randomness and reducing the likelihood of predictable word choices. Machine learning classification using Support Vector Machines (SVM) of 98% and Random Forests of 95.75% achieved high accuracy in differentiating between AI and human essays, highlighting the effectiveness of linguistic features for automated detection. The study concludes that AI-generated content can be reliably distinguished from human writing using stylistic and lexical features, contributing to the development of more reliable AI assessment tools and a better understanding of NLP model behavior.

General Terms
Pattern Recognition, Algorithms, AI-generated, Predictive Patterns

Keywords
Predictive Patterns, AI-generated essays, DeepSeek, ChatGPT, Gemini, Machine learning Analyzing word frequency

1. INTRODUCTION
Recent advances in natural language generation have significantly improved the diversity and quality of texts generated by chatbots such as ChatGPT and DeepSeek, making them almost indistinguishable from human-written texts. This has raised concerns about potential misuse, including the spread of misinformation or disruption of educational systems [1]. Research highlights that AI-generated texts often exhibit specific characteristics, such as repetitive phrasing and lack of depth in contextualization [2] [3] which poses a challenge for AI-generated content in academic contexts, creative writing and professional communication where thorough contextualization and critical engagement with existing literature are needed [4][5]. AI-generated text and machine generated text avoids language that is not commonly used and lack emotional semantics and personal biases found in human language [6].

Different writers, be they human or machine, exhibit distinct linguistic fingerprints. This perception raises questions about the underlying mechanisms of AI word prediction and the extent to which specific patterns can be attributed to AI-generated text. Studies have shown that repetitive phrasing and a lack of depth in contextualization have been attributed to the algorithms used in training AI models, which tend to prioritize coherence and clarity over understanding and originality [7] while for humans, these fingerprints are shaped by individual experiences, education, and emotional depth. Understanding these patterns is crucial for refining AI models and improving their applicability across different contexts, including education, creative writing, and professional communication.

The aim of artificial intelligence (AI) has been to develop intelligent systems capable of using language as proficiently as humans, facilitating fluent conversations and a meticulous comprehension of language intricacies. According to McShane and Nirenburg, language processing within AI models is conceptualized from an agent perspective, integrated into a broader model of perception, reasoning, and action [8]. Central to this perspective are the core prerequisites for success, including the ability to extract meaning from linguistic expressions, represent them in memory, and utilize these representations for decision-making across verbal, physical, and mental actions. The multifaceted nature of linguistic phenomena ranges from morphological ambiguity to pragmatic ambiguity. Semantic analysis emerges as a pivotal sub-task of NLP, enabling computers to derive meaning from textual data through grammatical analysis and contextual interpretation. Semantic classification models, including topic classification,

sentiment analysis, and intent classification, demonstrates the practical applications of semantic analysis in various domains, from customer service to marketing analytics. Linguistic analysis provides a rich theoretical framework and methodological insights crucial for understanding the complexities of language generation in AI systems [8].

The emergence of generative artificial intelligence (AI) challenges traditional notions of creativity, prompting a reevaluation of its essence and its relationship to human creativity [9]. Generative AI exhibits an uncanny ability to produce original content resembling human creative choices, such as writing, painting, and composing music, blurring the lines between human and machine creativity. Despite operating on algorithmic principles, generative AI derives its rules from training data, simulating human-like creative processes. Two responses have emerged in the creative sector: one suggesting that AI lacks individual expression characteristic of human creativity, while the other argues that AI merely recombines existing cultural elements into new forms, devoid of genuine creativity. The rise of generative AI challenges conventional notions of creativity, raising fundamental questions about its nature and the role of machines in creative endeavors [9].

The primary focus for detecting AI-generated text is linguistic analysis, which breaks out syntactic patterns, word choices, and sentence structures. When a person uses too many words, repeats the same thing, or breaks the rules, this is a red signal and anomaly detection methods point out when language patterns are broken. Machine learning models trained to spot anomalies can distinguish AI writing from human written language. Determining if writing was created by AI is complicated and ever-changing. Linguistic signals, inconsistency analysis, information inspection, stylometric quirks, bias identification, outliers, and purpose-built models are crucial [10].

In 2023, an analysis [11] revealed a notable and uneven surge in the frequency of certain keywords, both individually and collectively. It is speculated that a minimum of 60,000 publications (accounting for just above 1% of total articles) received assistance from LLMs. This figure could potentially be adjusted and further detailed through the examination of additional paper attributes or the discovery of more keywords suggestive of LLM involvement.

[12] study investigates the impact of Artificial Intelligence (AI),specifically generative AI technologies (GAI), on the linguistics of academic article titles. Triggered by suspicious of increased usage of specific verbs in article titles, this research hypothesizes that GAI tools may be influencing the language of scientific communication. To explore this hypothesis, we conducted a comprehensive analysis on the frequency and distribution of 15 selected verbs in research article titles, using data extracted from the SCOPUS database spanning 2015 to 2024. The methodology integrates qualitative observations with a bibliometric approach, examining the presence and trends of these verbs across multiple scientific disciplines. The findings reveal a marked increase in these verbs, pointing towards AI's involvement in title generation. We also explore document characteristics, such as disciplinary backgrounds and publication contexts, to gauge AI's impact on academic writing. Furthermore, the research attempts to quantify the extent of AI-assisted title generation. Despite several limitations, this investigation paves the way for future studies to broaden the linguistic and database scope. It underscores the need for establishing AI usage standards in academic publishing, contributing valuable insights into the ongoing dialogue about AI's integration into academic writing.

The quick development of AI-generated writing has sparked serious questions about linguistic distinctiveness, plagiarism, and validity. Differentiating between machine-generated and human-authored material has become a crucial difficulty as AI models like as ChatGPT, DeepSeek, and Gemini create content that seems more and more human. Because students and content producers may abuse AI systems to produce essays, reports, or articles without giving due credit, publishers and educators are concerned about the hazards of plagiarism. The diversity and originality of written speech are threatened by the lack of linguistic distinctiveness in AI outputs, which are frequently typified by repeated phrasing, predictable word selections, and generic structures. These worries also extend to the ethical and legal spheres, where AI-generated impersonation or false information may erode confidence in digital interactions.

This research aims to explore the predictive mechanisms underlying word choice in AI systems, investigate the parameters that influence linguistic outputs, and analyze the frequency of commonly used words to identify potential markers of AI-generated content.

2. LITERATURE REVIEW

2.1 Text Representation

In natural language processing (NLP), text representation is a fundamental step that involves converting textual data into formats that machine learning models can process effectively (Worth, 2023). This typically encompasses two primary stages: tokenization and word embedding.

2.2 Tokenization

Tokenization is the process of dividing text into smaller units known as tokens, which can be words, subwords, or characters. Recent studies have highlighted the significance of tokenization strategies in handling morphologically rich and low-resource languages. For instance, MorphTok introduces a morphology-aware segmentation approach that improves tokenization for Indian languages by incorporating linguistic features into the tokenization process [13]. Similarly, research by [14] demonstrates that sentence piece tokenization outperforms Byte Pair Encoding (BPE) in zero-shot Named Entity Recognition tasks for Indic languages, owing to its better preservation of linguistic structures. Furthermore, [15] propose an optimized BPE configuration that reduces token counts and enhances performance, particularly in low-resource language models.

2.3 Word Embedding

Word embeddings are numerical representations of words in a continuous vector space, capturing semantic relationships based on contextual usage. Advancements in this area include the development of sense-aware contextualized word embeddings, which effectively encode semantic changes over time and context [16]. Additionally, [17] discusses the evolution of word embedding techniques and their applications in capturing semantic spaces in NLP. In specialized domains, such as psychiatric speech analysis, word embeddings have been utilized to unravel the structure of meaning in psychosis, demonstrating their versatility and applicability across various fields [18].

Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 8 – No. 1, January 2026 – www.caeaccess.org

Figure 2.1: Word embeddings projection from 60 dimensions into 2 dimensions, similar words are clustered close to each other in semantic space, blue color consider neutral words, red color negative words and green positive.

Words such as *"bad"*, *"worst"*, *"good"*, and *"nice"* can be analyzed within a semantic space to understand their emotional and contextual similarities. For example, *"bad"* and *"worst"* are semantically related through their negative connotations, whereas *"good"* and *"nice"* are associated with positive sentiment and are semantically distant from the negative ones [18]. In a semantic vector space, words with similar meanings cluster close together, while words with different meanings are positioned farther apart, revealing underlying linguistic relationships [16].

Word embeddings used in modern language models are dense vectors—continuous numerical representations without zero elements—that typically range between 50 and 1000 dimensions [19]. These embeddings are generated through self-supervised training on large-scale corpora, enabling models to automatically learn contextual and semantic patterns without manually labeled data [20]. Among the popular word embedding techniques, Word2Vec represents a static approach where each word is assigned a fixed vector regardless of context. For instance, in the sentences *"We trained a deep neural network to recognize patterns in the data"* and *"We designed a network to enhance the data transfer rate"*, the term *"network"* would be encoded identically despite its distinct meanings—one referring to a machine learning structure, and the other to a communication system [20].
In contrast, contextual embeddings such as those generated by BERT (Bidirectional Encoder Representations from Transformers) consider the surrounding context of a word. Here, the model distinguishes between different meanings of *"network"* and assigns different vectors accordingly. For example, in the phrase *"neural network"*, the embedding for *"network"* would be close to other AI-related terms, reflecting its usage within the domain of deep learning [16][19]. There is also sparse embedding, where the vector contains zero elements. One widely used method to generate sparse embeddings is TF-IDF (Term Frequency-Inverse Document Frequency), which evaluates the importance of a term within a document relative to a collection of documents. The process typically involves three stages: computing term frequency, computing inverse document frequency, and multiplying the two to generate a weighted representation [20].

First stage is counting the number of times term t occurs in document d in some training corpus, and then taking the log of the count plus 1 to avoid undefined behavior when the count is 0, the equation is given by:

$$TF(t, d) = \log_{10}(\text{count}(t, d) + 1) \quad (1)$$

Second stage is to count the number of document DF in which contains the term t, then calculating the IDF as:

$$IDF = \log_{10}(N/DF) \quad (2)$$

where N is the total number of documents in training corpus. Third stage is to calculate the weighting TF-IDF as:

$$W_{t,d} = TF_{t,d} * IDF_t \quad (3)$$

The limitation of this method is the dimension of the vectors representing words grows quickly e.g., if there is |V|=1000000 unique words in corpus then the dimension is |V|, also this methods dose not capture the semantic, context and other structures of words and sentences.

2.4 ChatGPT and GPT-3.5 Series

ChatGPT is a large language model based on the Generative Pre-trained Transformer (GPT) architecture, developed by OpenAI. It currently exists in two versions: GPT-3.5 and GPT-4, with GPT-4 being the most recent and advanced version ([21]. The GPT-3.5 series is an improvement over the original GPT-3 models and is designed for general-purpose language generation. These models can generate coherent text in response to a wide range of prompts and instructions [23].

Among the prominent models in the GPT-3.5 series are text-davinci-003 and gpt-3.5-turbo. While text-davinci-003 is effective for text completion tasks, gpt-3.5-turbo being a fine-tuned and optimized version excels in chat-based applications and offers better performance at a significantly lower cost (approximately one-tenth) compared to text-davinci-003 when accessed via OpenAI's API [22].

2.5 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a powerful transformer-based language model that was pre-trained on massive corpora, including the BooksCorpus (800 million words) and English Wikipedia (2.5 billion words). It uses the WordPiece tokenization method with a vocabulary of approximately 30,000 subwords to convert text sequences into token embeddings [18] [23]. BERT introduces special tokens such as [CLS] to indicate the start of a sequence and [SEP] to mark the end. The embedding of the [CLS] token is commonly used to represent the entire input sequence for downstream tasks like sentence classification (Wang et al., 2023). There are two main versions of BERT: BERT Base (110 million parameters) and BERT Large (340 million parameters), with the latter offering improved performance on large-scale NLP benchmarks [24].

2.6 Machine Learning Classifiers

In the context of linguistic analysis and predictive pattern recognition, supervised machine learning classifiers have been widely adopted to automate text classification and evaluate language patterns [30]. These models learn to associate textual features with specific outcomes, making them highly effective for tasks such as author attribution, sentiment analysis, and distinguishing between AI-generated and human-authored texts [26]. Recent studies have demonstrated the utility of supervised classifiers in analyzing language structures, token distributions, and predictive trends across various types of generated content [27]. This section reviews three foundational classifiers—Logistic Regression, Support Vector Machines (SVM), and Random Forests—highlighting their roles, mechanisms, and relevance to analyzing linguistic and predictive patterns in AI-generated essays [28][29].

Supervised machine learning classifiers are algorithms trained on labeled datasets to predict outcomes based on input features. Some of the widely used classifiers are Logistic Regression, Support Vector Machines (SVM), and Random Forests.

2.6.1 Logistic Regression

Logistic Regression is a statistical model employed for binary classification tasks, estimating the probability that a given input belongs to a particular category. It utilizes the sigmoid function to map real-valued inputs into a range between 0 and 1, representing probability scores . Predictions are made by applying a threshold (commonly 0.5) to these probabilities; values above the threshold are classified into one category, and those below into another . Recent advancements have introduced dynamic logistic ensemble models, enhancing classification accuracy by recursively partitioning datasets and constructing multiple logistic models.

2.6.2 Support Vector Machine (SVM)

Support Vector Machines are supervised learning models used for classification tasks, aiming to find the optimal hyperplane that separates data points of different classes with the maximum margin . For datasets that are not linearly separable, SVMs employ the "kernel trick" to transform data into higher-dimensional spaces where a linear separator can be found . Recent research has proposed novel kernel functions, such as the Cholesky kernel, which consider the variance-covariance structure of the data to improve classification performance.

2.6.3 Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of their predictions for classification tasks . Each tree is trained on a random subset of the data and features, introducing diversity and reducing overfitting . Recent studies have focused on improving Random Forest algorithms by enhancing the accuracy and diversity of the individual trees, leading to better overall performance.

3. METHODOLOGY

3.1 Data Collection

The data collection process for this study involved collection of AI-generated and human-authored essays, with careful attention to variability in content style, topic diversity, and text generation parameters. For the AI-generated essay corpus, 1000 essays were created using three generative language models: ChatGPT, DeepSeek, and Gemini. To ensure balanced representation and comparability across models, approximately 333 essays were generated by each system with Gemini getting additional 1 essay to make it 334. The essay prompts covered a wide range of topics, including education, technology, social issues, and abstract concepts.

The human-authored essay corpus consist of a comparable set of essays compiled to serve as a benchmark for analysis. These essays were sourced from publicly available educational repositories, academic forums, student writing samples, and open-access essay databases. The selected essays were carefully matched in terms of topic and style with those generated by AI, to ensure consistency in content domain and facilitate direct comparison.

The dual-sourced datasets comprising of both machine-generated and human-written texts provided the foundation for analyzing word frequency distributions, next-word prediction behavior, and model-specific language patterns, in alignment with the core objectives of the study.

3.2 Linguistic Analysis: Next-Word Prediction Analysis

This analysis help evaluate how generative language models predict subsequent words in a text sequence, and how variables such as temperature settings, contextual cues, and token probability distributions influence these predictions. The analysis helps in understanding the decision-making process of AI models in sequence generation and lexical variation.

Let a sequence of tokens be represented as:

$$X = (x_1, x_2, x_3 \dots, x_t) \quad (4)$$

The task of next-word prediction involves estimating the probability distribution over the vocabulary V for the next token x_(t+1), given the context tokens:

$$P(x_{t+1}|x_1, x_2, x_3 \dots, x_t) = P(x_{t+1}|X) \quad (5)$$

A language model f such as GPT or BERT-style decoder) computes this distribution using:

$$P(x_{t+1}|X) = softmax(z) \quad (6)$$

Where $z \in R^{|V|}$ is the output logits vector produced by the model for the next word, and the softmax function is defined as:

$$softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{|V|} e^{z_j}} \quad (7)$$

Each corresponds to the log it score of the i-th vocabulary token.

3.2.1 Temperature Scaling

Temperature τ is a hyperparameter that controls the "creativity" or randomness of the output distribution. Temperature scaling modifies the log its before applying softmax:

$$P(x_{t+1} = w_i|X) = \frac{e^{z_{i/\tau}}}{\sum_{j=1}^{|V|} e^{z_{j/\tau}}} \quad (8)$$

τ<1: Sharper distribution (model becomes more confident and deterministic).

τ>1: Smoother distribution (more diverse and creative outputs).

We conduct this experiment at three temperature levels: 0.2 (low randomness), 0.7 (moderate randomness), and 1.0 (high randomness).

3.2.2 Contextual Cue Evaluation

To assess the impact of context, the conditional probability $P(x_{t+1}|X)$ is measured under two conditions:

Full context: the complete sequence $x_1, x_2, x_3 \dots, x_t$

Reduced context: truncated or noisy context (remove preceding sentence or mask key tokens)

The contextual influence score C can be defined as the Kullback-Leibler (KL) divergence between the next-word distributions under full and reduced contexts:

$$C = D\,(P_{full} || P_{reduce}) = \sum_{i=1}^{|V|} P_{full}(w_i) \log {P_{full}(w_i)}/{P_{reduce}(w_i)} \quad (9)$$

This measures how much the context changes the next-token prediction.

3.2.3. Token Probability Analysis

For each generated word x_{t+1}, we record:

The maximum predicted probability: $max_i\ P(x_{t+1} = w_i|X)$

The rank of the selected token in the vocabulary distribution

The entropy of the output distribution:

$$H(X) = -\sum_{j=1}^{|V|} P(x_{t+1} = w_i|X)\ log\ P(x_{t+1} = w_i|X) \quad (10)$$

Entropy measures the uncertainty of the model at each prediction step. Higher entropy suggests more uncertainty and creativity; lower entropy indicates more deterministic output.

3.3 Analytical Techniques: Word Frequency Analysis

To identify distinctive lexical patterns between human and AI-generated texts, we employed TF-IDF (Term Frequency-Inverse Document Frequency) to evaluate the importance of words in the corpus, emphasizing words unique or particularly significant to each class. Additionally, n-grams analysis (bigrams, trigrams, etc.) was utilized to capture contextual word sequences and phrases that may differentiate writing styles.

3.4 Predictive Pattern Analysis

Analyzing next-word probability distributions enabled us to examine the fluency and coherence patterns inherent in human versus AI texts. By modeling the likelihood of subsequent words given previous contexts, we identified patterns indicative of AI-generated text, which often exhibit repetitive or predictable sequences. This approach provided insight into the structural differences in language generation, supporting feature engineering for classification models.

3.5 Classification Testing

We conducted machine learning classification experiments to evaluate the ability to distinguish AI-generated essays from human-written ones. Algorithms such as Support Vector Machines (SVM) and Random Forests were trained on features derived from frequency analysis and predictive pattern metrics. Model performance was assessed using standard metrics like accuracy, precision, recall, and F1-score, along with confusion matrices to analyze classification errors and model robustness.

3.6 Evaluation metrics

This section describes the evaluation metrics used for evaluating the methods for differentiating between AI-generated essay and human written essay.

3.6.1 Precision, Recall, Accuracy and F1 score

Precision, recall and F1 are metrics that measure how well a classifier performs on predicting the correct target class. The metrices are derived from confusion matrix which is for binary classification problem a 2×2 matrix, where the rows are the true classes, and the columns are the predicted class. The upper left cell in confusion matrix is called true negative (TN) which contains the instances of negative class that has been correctly classified as negative, the lower right cell is true positive (TP) which is the target class in interest and contains the instances that has been correctly classified as positive, the lower left cell is false negative (FN) cell and contains the instances that has been incorrectly classified as negative, The last cell is the upper right cell called false positive (FP) that contains the instances that has been misclassified as positive while they belong to the negative class. These metrics are calculated based on the confusion matrix as follows:

Accuracy: is the ratio of correctly classified instances out of the total instances, this metric is not suitable for imbalanced dataset and dose not gives how well the classifier perform on specific class, but it gives the performance for classifying both negative and positive classes.

$$\text{Accuracy}=(TP+TN)/(TP+TN+FP+FN) \quad (11)$$

Precision: is the ratio of true positive class out of all positive prediction. Precision measures the quality of the classifier when it predicts the positive class where high precision indicates a low rate of false positive errors, but precision does not give how much of all positive classes the classifier could identify therefore precision is combined with another metrics such Recall.

$$\text{Precision}=TP/(TP+FP) \quad (12)$$

Recall: is the ratio of true positives out of all actual positive instances, it measures how well the classifier identifies positive instances, where high recall indicates a low rate of false negative errors, recall is often combined with precision for more precise measurement.

$$\text{Recall}=TP/(TP+FN) \quad (13)$$

F1 Score: this metric is called the harmonic mean of both precision and recall, this combines precision and recall into a one single metric that evaluates the trade-off between precision and recall, it can be used to compare classifiers. The value of F1 score is high if both precision and recall are high while if precision or recall has low value then F1 score will be low.

$$F1=2/(1/PPPrecision+1/Recall) = 2*(Precision* Recall)/ (Precision+Reall) \quad (14)$$

4. RESULTS AND DISCUSSIONS

Next-Word Prediction Analysis

The results in table 1 and figure 2 demonstrates that temperature settings significantly influence next-word prediction probabilities. Using a sample sentence "In the digital age social media has become an inte...", the prediction probabilities for the word "res" at temperature settings of 0.7, 1.0, and 1.5 are 0.8563, 0.4896, and 0.0763, respectively. As the temperature increases from 0.7 to 1.5, the prediction probability of the word "res" decreases dramatically. The probabilities of other words such as "ress," "view," "get" also change with temperature settings. This indicates that higher temperature values increase the randomness of word selection, reducing the likelihood of highly probable words being chosen. This aligns with the research objective of examining how NLP parameters like temperature influence word selection.

Table 1. Summary Table of Predictions (Temperature vs Word Probability)

	Temperature	Predicted Word	Probability
0	0.7	Res	0.856321
1	0.7	Ress	0.05806
2	0.7	View	0.026478
3	0.7	-	0.009751

4	0.7	Get	0.005571
5	1	Res	0.489584
6	1	Ress	0.074421
7	1	View	0.042954
8	1	-	0.021346
9	1	Get	0.014426

10	1.5	Res	0.076271
11	1.5	Ress	0.021724
12	1.5	View	0.015059
13	1.5	-	0.009448
14	1.5	Get	0.007276

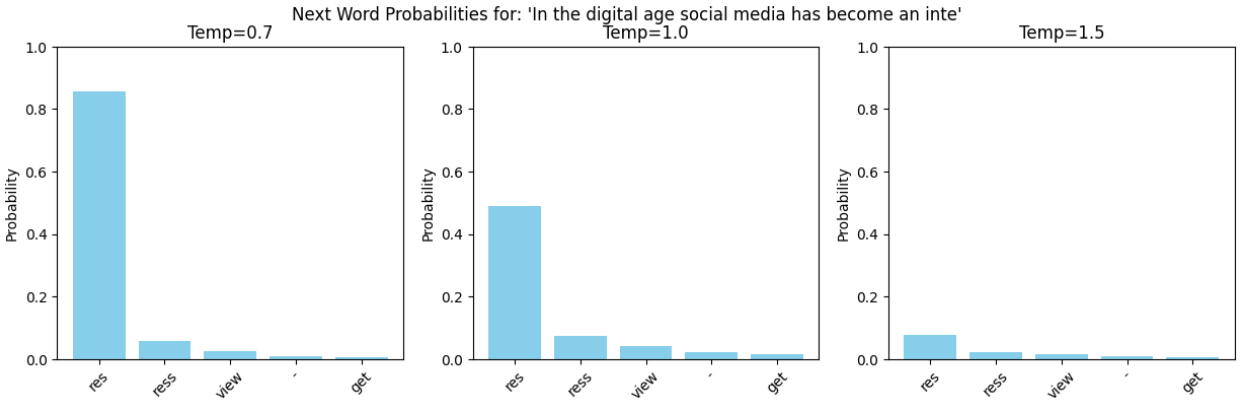

Fig 2: Temperature probability of Next-word prediction for essay: 'In the digital age social media has become an inte...'

Word Frequency Analysis using TF-IDF (Term Frequency–Inverse Document Frequency)

Word frequency analyses using TF-IDF revealed distinctive vocabularies. Figure 3 and figure 4 shows the word frequency analysis applied TF-IDF to distinguish common words in AI-generated vs. human-authored essays. The AI-generated texts often contain words related to cooperation, media, education, and technology, such as "children," "media," "cooperation," "famous," and "skills." The human essays tend towards personal, social, and experiential words like "like," "think," "know," "friends," "love," "school," etc. The Bigrams highlight these differences of AI texts frequently include technical and systematic phrases ("childrenlearn," "helpchildren," "longterm," "educationlife"), whereas human texts show colloquial and social expressions ("feel like," "think going," "good thing"). Table 2 shows the top typical words frequency analysis using TF-IDF.

Table 2: Top AI/Human

Top AI-typical words (TF-IDF)	['children' 'media' 'cooperation' 'famous' 'learn' 'cooperate' 'education' 'competition' 'privacy' 'ai' 'skills' 'celebrities' 'compete' 'learning' 'games' 'public' 'important' 'data' 'systems' 'taught']
Top Human-typical words (TF-IDF)	['really' 'just' 'don' 'know' 'like' 'think' 'going' 'want' 'feel' 'good' 'time' 'wonder' 'right' 'friends' 'love' 'guess' 've' 'home' 'minutes' 'class']

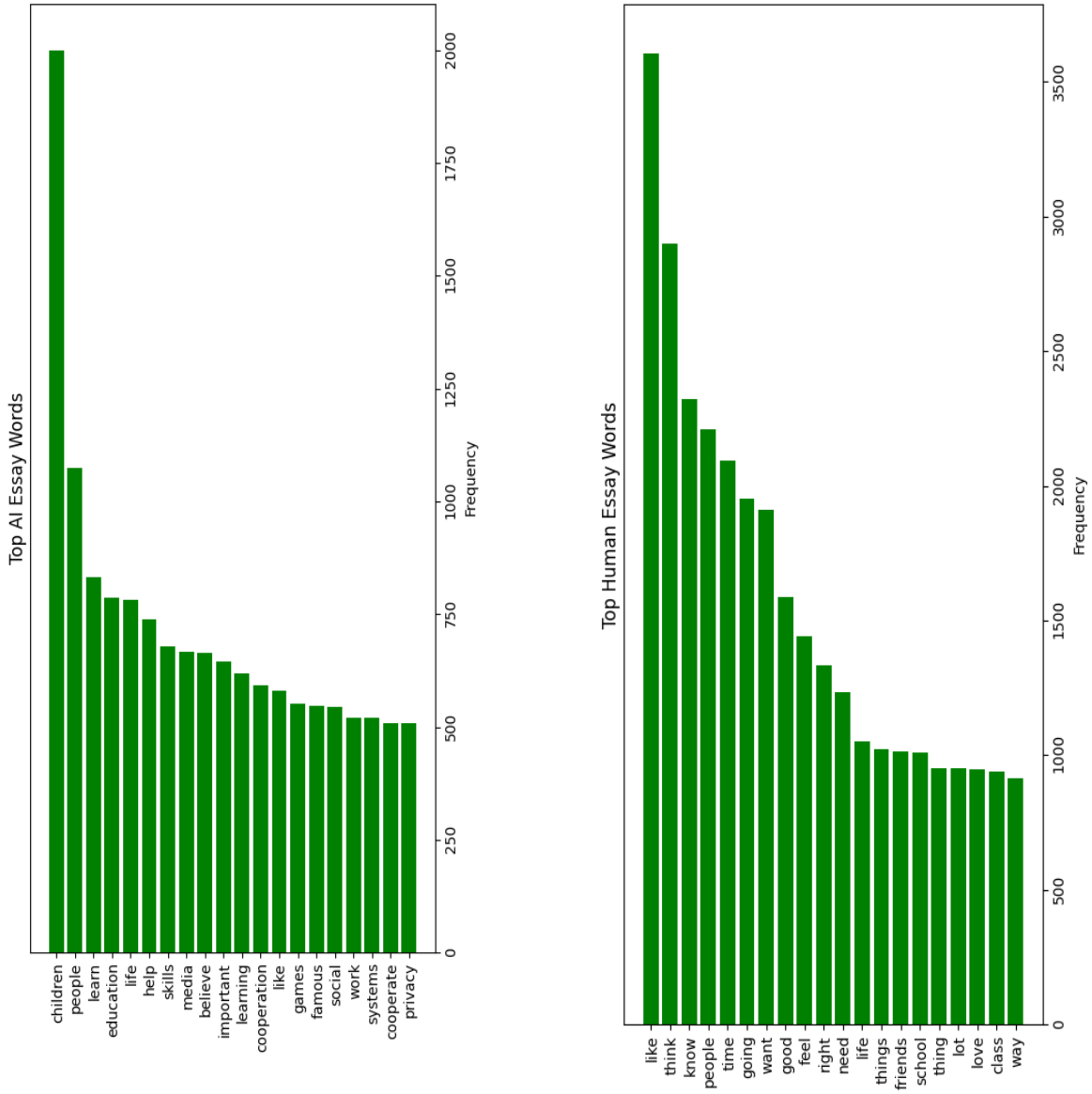

Fig 3: Distinguish common words in AI-generated vs. human-authored essays

Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 8 – No. 1, January 2026 – www.caeaccess.org

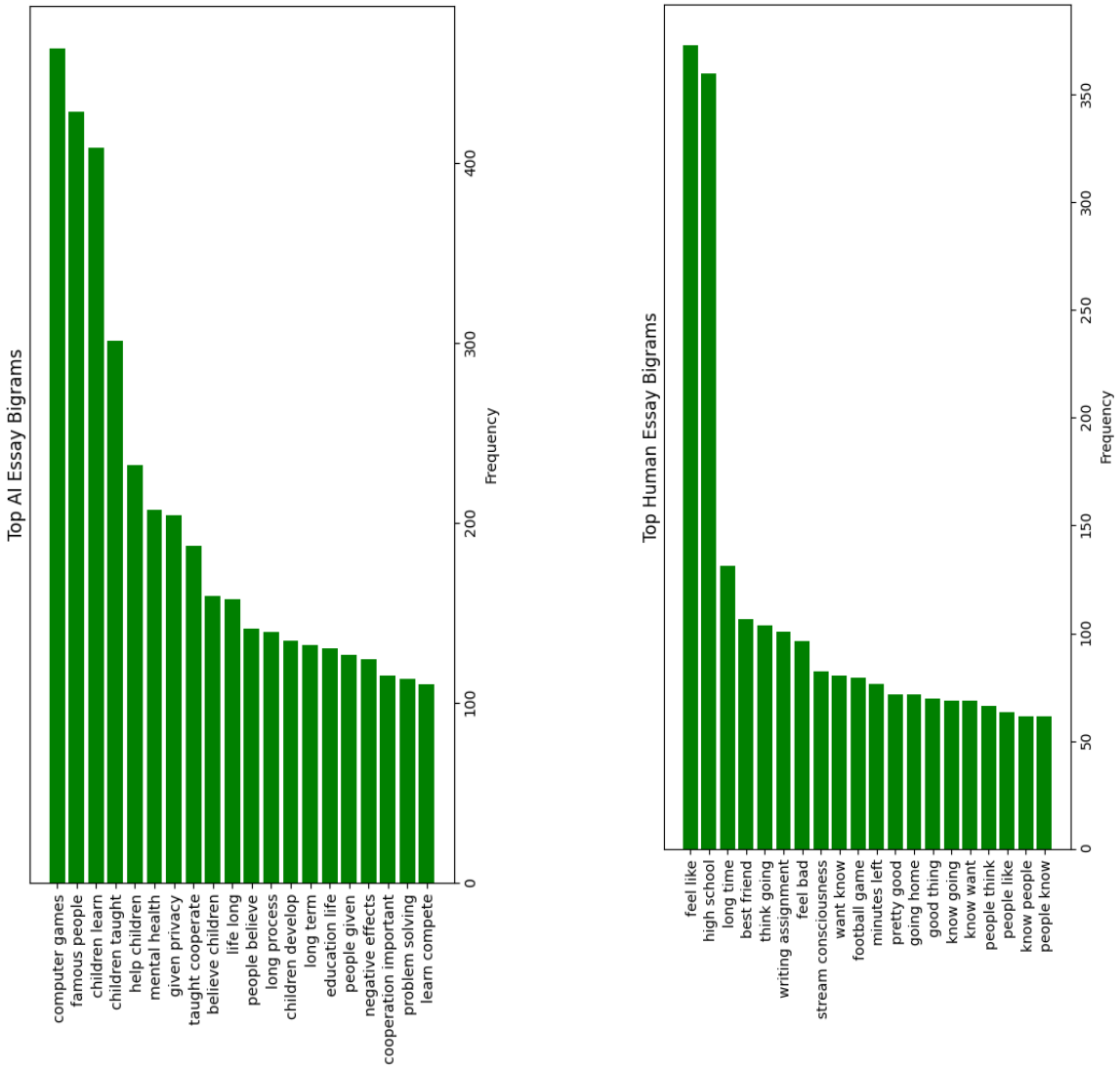

Figure 4: Top AI-generated Bigrams and human-authored essays Bigrams

Figure 5 shows the bar chart displaying the TF-IDF score difference for top words between AI-generated and human-authored essays. The x-axis represents the TF-IDF score difference, and the y-axis lists the words. The chart highlights words such as "children," "media," and "cooperation" as more distinctive in AI-generated essays.

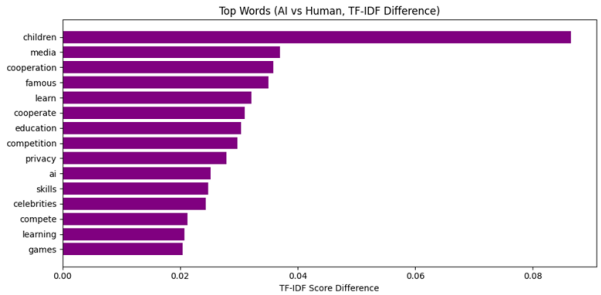

Figure 5: TF-IDF score difference for top words between AI-generated and human-authored essays

The bar charts in figure 6a, 6b and 6c shows the frequency of top bigrams in Gemini-generated, Deepseek-generated and ChatGPT-generated essays. The x-axes represents bigram frequency, and the y-axes lists the bigrams. The chart highlights bigrams like "long term," "older adults," and "public health" as frequent in Gemini-generated essays, bigrams such as "computer games," "help children," and "mental health" as frequent in Deepseek-generated essays and bigrams like "children learn," "famous people," and "children taught" as frequent in ChatGPT-generated essays.

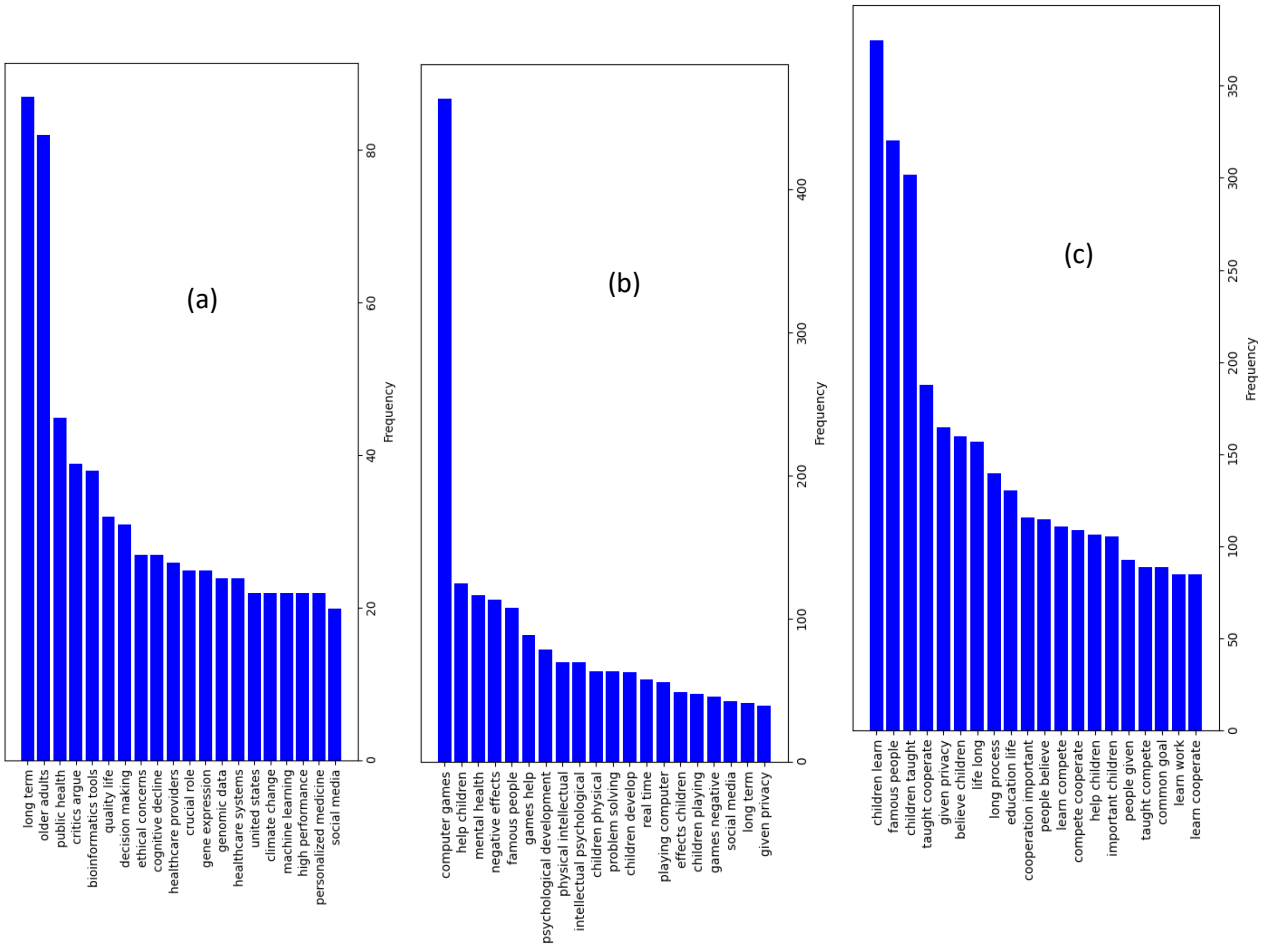

Figure 6: Top Gemini(a), DeepSeek(b) and ChatGPT(c) generated essays Bigrams

Figure 7(a) shows a scatter plot representing the clustering of AI-generated and human-authored essays. The x-axis represents PCA Component 1, and the y-axis represents PCA Component 2. The chart illustrates the separation between AI and human essays in the PCA space. The color of the points is used to distinguish between AI-generated essays (yellow) and human-generated essays (purple). The distribution of points shows a clear separation between the two groups, indicating that the essays can be distinguished based on the features extracted. The clustering pattern suggests that there are distinct stylistic or content-related characteristics that differentiate AI-generated texts from those produced by humans.

Figure 7(b) presents a three-dimensional Principal Component Analysis (PCA) visualization of Term Frequency-Inverse Document Frequency (TF-IDF) vectors, contextualized by cosine similarity. This graphical representation is instrumental in understanding the distribution and clustering of AI-generated and human-generated text data. The color coding is used to differentiate between AI-generated texts (blue) and human-generated texts (red), providing a visual distinction between the two categories. The clusters of points are indicative of the cosine similarity context, where texts with similar content and structure are grouped together. The visualization reveals that while there is a general overlap between AI and human-generated texts, there are distinct regions where one category dominates over the other. This suggests that certain characteristics of the texts can be differentiated based on their origin. The clear demarcation of clusters also suggests potential areas for improvement in AI text generation to bridge the gap with human-like text characteristics.

Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 8 – No. 1, January 2026 – www.caeaccess.org

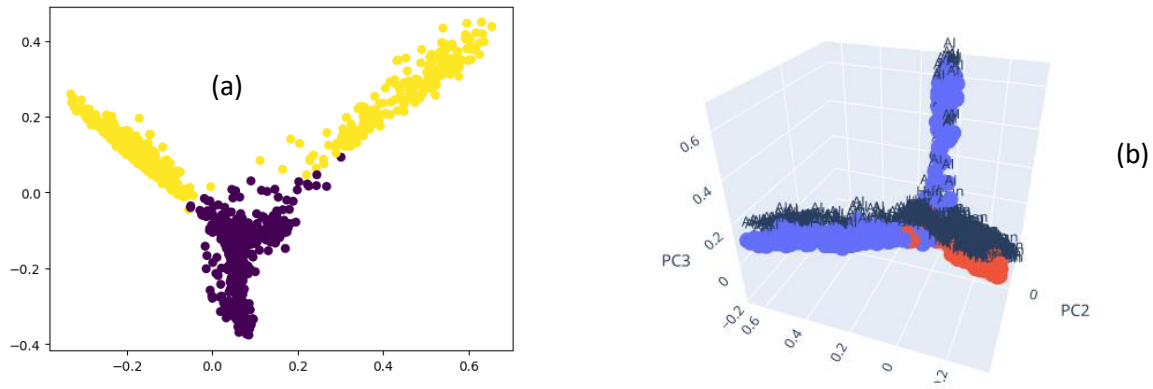

Figure 7: Figure 7(a) shows the scatter plot representing the clusterings of AI-generated and human-authored essays while figure 7(b) shows the 3D PCA of TF-IDF Vectors with cosine similarity context.

Machine Learning classification

t-Distributed Stochastic Neighbor Embedding (t-SNE) Mapping of AI vs Human Essays

The t-SNE in figure 8 shows a two-dimensional representation of high-dimensional essay data, distinguishing between AI-generated (blue) and human-authored (red) texts. In the scatterplot, each point represents an individual essay, and their spatial arrangement is informed by their linguistic and semantic similarities. The result reveals a clear clustering pattern. AI-generated essays form distinct, tight clusters, especially on the right side of the plot, whereas human-written essays are mostly concentrated on the left, though with a more dispersed structure. This separation strongly suggests that AI and human essays exhibit consistent, distinguishable linguistic patterns, validating the premise that AI-generated text can be identified through its structural and lexical characteristics.

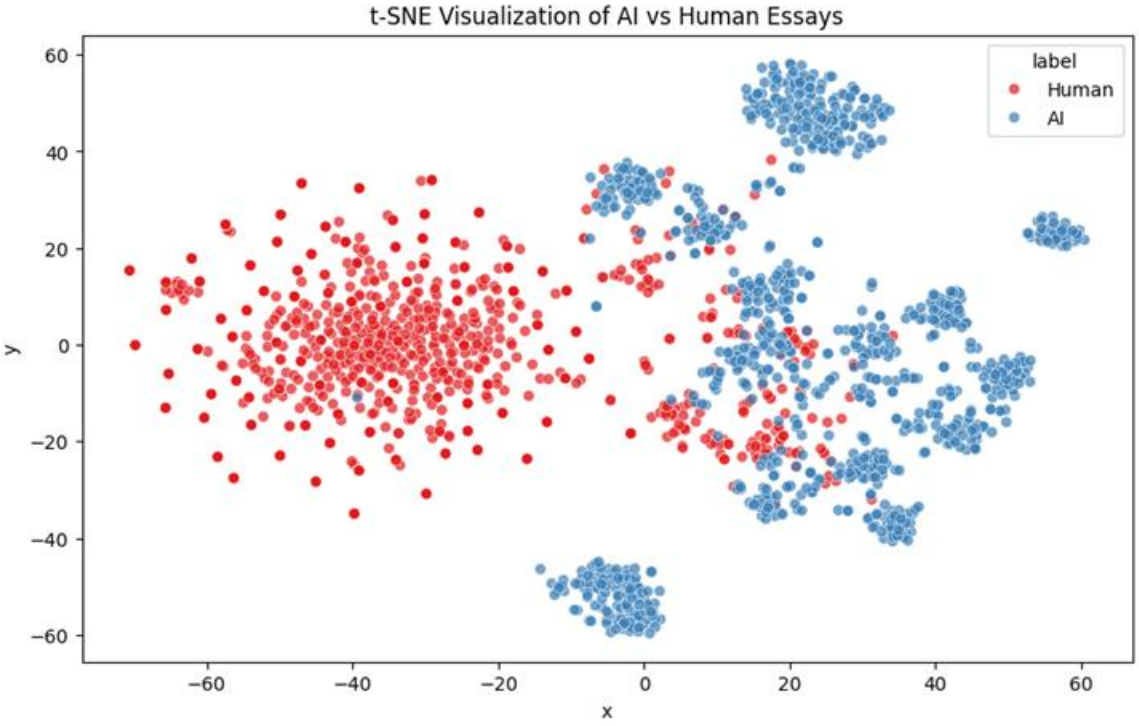

Figure 8: t-SNE Visualization of AI and Human Essays.

Classification Performance Evaluation

The classification models used to differentiate between AI-generated and human-written essays demonstrated remarkably high accuracy, reinforcing the clear separation previously observed in the t-SNE plot. As shown in Table 3, the first classifier achieved an overall accuracy of 98%. It recorded a precision of 1.00 for human essays and 0.97 for AI essays, with recall values of 0.97 for human essays and 1.00 for AI essays. The average F1-score of 0.98 indicated a strong balance between precision and recall for both categories. Similarly, the Random Forest classifier also performed well, attaining an accuracy of 95.75%. Its confusion matrix showed that all AI essays were correctly identified, although 17 human essays were misclassified as AI. Both classes achieved F1-scores of 0.96, demonstrating the model's robustness. These findings underscore the effectiveness of linguistic features in distinguishing between AI and human texts, suggesting they

are highly discriminative and well-suited for automated detection. Notably, the perfect recall for AI essays across both models indicates that AI-generated content often exhibits repetitive or formulaic patterns that are easily and consistently detected by the classifiers.

Table 3: Results of the two classification models used for both AI-generated and Human essays

		SVM				RANDOM FOREST			
		precision	Recall	f1-score	support	precision	recall	f1-score	support
	0	1	0.97	0.98	199	1	0.91	0.96	199
	1	0.97	1	0.99	201	0.92	1	0.96	201
accu	racy			0.98	400			0.96	400
macro	avg	0.99	0.98	0.98	400	0.96	0.96	0.96	400
weighted	avg	0.99	0.98	0.98	400	0.96	0.96	0.96	400

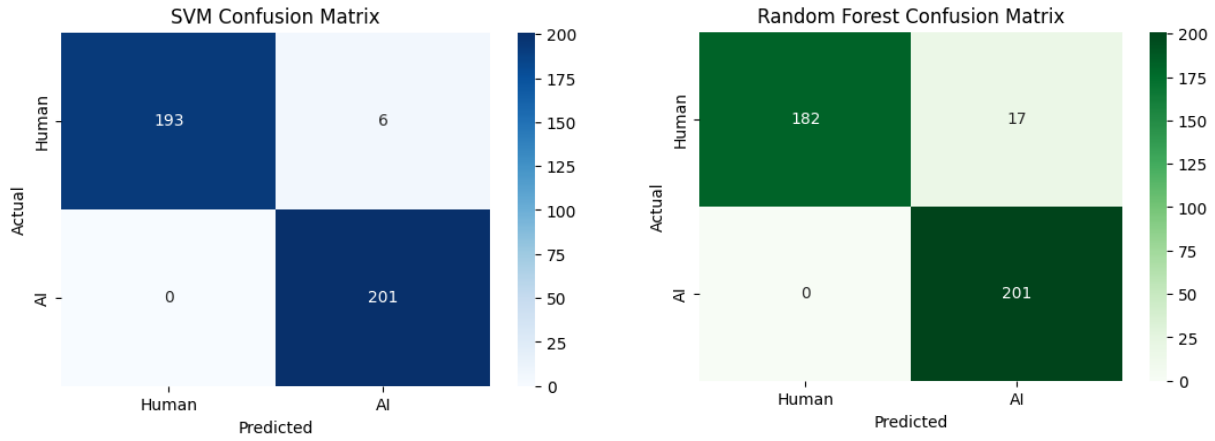

Figure 9: Confusion Matrix for both SVM and Random Forest

The bar chart displaying the top 20 important features (as determined by a Random Forest model) highlights the words that most significantly contribute to distinguishing between AI-generated and human-authored essays. The importance scores, which range from approximately 0.01 to 0.06, indicate the relative influence of each feature in the classification process. Words such as "don," "going," and "really" are among the most influential features, suggesting that these words are more indicative of either human or AI authorship. This insight is crucial for understanding the linguistic cues that machine learning models use to differentiate between the two types of essays.

The ROC curve, which shows the performance of both SVM and Random Forest models, indicates that both models have excellent classification capabilities, with an AUC (Area Under the Curve) of 1.00. This suggests that the models can perfectly distinguish between AI and human essays based on the features extracted. The high true positive rate and low false positive rate demonstrate the models' robustness in classifying essays accurately.

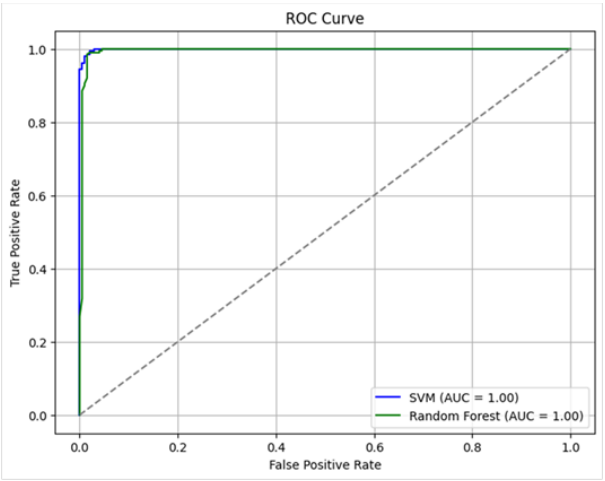

Figure 10: ROC Curve for SVM and Random Forest

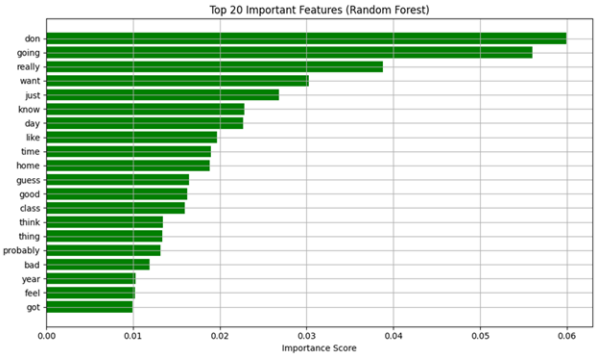

Figure 11: Top 20 RF Model important features to distinguish between AI-generated and human-authored essays Curve

5. CONCLUSION

This research contribute to a deeper understanding of the linguistic characteristics of AI-generated content, with implications for improving AI systems and addressing misconceptions about AI writing. These results suggest that it is possible to reliably differentiate between human and AI essays using machine learning models based on stylistic and lexical features. The results indicate that AI-generated essays tend to use a distinct vocabulary, when compare to human-authored essays. The analysis of next-word prediction algorithms revealed that temperature settings significantly influence word selection probabilities in AI models. Higher temperature values increase the randomness of word choice, reducing the likelihood of selecting highly probable words. The machine learning classification using SVM and Random Forests achieved remarkable accuracy in differentiating between AI and human essays, reinforcing the clear separation observed in linguistic patterns. In conclusion, this research has contributed to a deeper understanding of the linguistic characteristics of AI-generated content, with implications for improving AI systems and addressing misconceptions about AI writing.

Future research can extend this work by incorporating deeper contextual and semantic features, as well as transformer-based models, to improve robustness across diverse genres, languages, and evolving AI systems. Additionally, longitudinal studies can explore how advancements in generative models and adaptive temperature controls influence detectability over time and inform fair, ethical AI-detection frameworks.

6. REFERENCES

[1] Tang, R., Chuang, Y. N., & Hu, X. (2024). The science of detecting LLM-generated text. *Communications of the ACM*, *67*(4), 50-59.

[2] Logacheva, E., Hellas, A., Prather, J., Sarsa, S., & Leinonen, J. (2024). Evaluating Contextually Personalized Programming Exercises Created with Generative AI. arXiv preprint arXiv:2407.11994. https://doi.org/10.1145/3632620.3671103

[3] Javaid, M., Haleem, A., Singh, R. P., Khan, S., & Khan, I. H. (2023). Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system. BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 3(2), 100115. https://doi.org/10.1016/j.tbench.2023.100115

[4] Draxler, F., Werner, A., Lehmann, F., Hoppe, M., Schmidt, A., Buschek, D., & Welsch, R. (2024). The AI ghostwriter effect: When users do not perceive ownership of AI-generated text but self-declare as authors. ACM Transactions on Computer-Human Interaction, 31(2), 1-40. https://doi.org/10.1145/3637875

[5] Dergaa, I., Chamari, K., Zmijewski, P., & Saad, H. B. (2023). Fromhuman writing to artificial intelligence generated text: Examiningthe prospects and potential threats of ChatGPT in academic writ-ing. Biology of Sport, 40(2), 615–622

[6] Roberto, C., & Sebastian, L. A. *One-Class Learning for AI-Generated Essay Detection (2023)*. : Corizzo, R.; Leal-Arenas, S. One-Class Learning for AI-Generated Essay Detection. Appl. Sci. 2023, 13, 7901. Hz

[7] Melliti, M. (2024). Using Genre Analysis to Detect AI-Generated Academic Texts. *Diá-logos*, *16*(29), 09-27.

[8] Akinwande, M., Adeliyi, O., & Yussuph, T. (2024). Decoding AI and Human Authorship: Nuances Revealed Through NLP and Statistical Analysis. International Journal of Cybernetics and Informatics. Vol. 13(4): 85-103

[9] Moreno A. and Redondo T. (2016). Text Analytics: the convergence of Big Data and Artificial Intelligence. IJIMAI 3, 6 (2016), 57–64.

[10] Shah, A., Ranka, P., Dedhia, U., Prasad, S., Muni, S., & Bhowmick, K. (2023). Detecting and unmasking AI-generated texts through explainable artificial intelligence using stylistic features. *International Journal of Advanced Computer Science and Applications*, *14*(10) 1043-1053

[11] Gray, A. (2024). ChatGPT" contamination": estimating the prevalence ofLLMs in the s cholarly literature. arXiv preprint arXiv. 2403.16887

[12] Comas-Forgas, R., Koulouris, A., & Kouis, D. (2025). 'AI-navigating'or 'AI-sinking'? An analysis of verbs in research articles titles suspicious of containing AI-generated/assisted content. *Learned Publishing*, *38*(1), 1-11.

[13] Brahma, M., Karthika, N. J., Singh, A., Adiga, D., Bhate, S., Ramakrishnan, G., Saluja, R., & Desarkar, M. S. (2025). MorphTok: Morphologically Grounded Tokenization for Indian Languages. *arXiv preprint arXiv:2504.10335*. https://doi.org/10.48550/arXiv.2504.10335

[14] Pattnayak, P., Patel, H. L., & Agarwal, A. (2025). Tokenization Matters: Improving Zero-Shot NER for Indic Languages. *arXiv preprint arXiv:2504.16977*. https://doi.org/10.48550/arXiv.2504.16977

[15] Raj, B. S., Suri, G., Dewangan, V., & Sonavane, R. (2024). When Every Token Counts: Optimal Segmentation for Low-Resource Language Models. *arXiv preprint arXiv:2412.06926*. https://doi.org/10.48550/arXiv.2412.06926

[16] Aida, T., & Bollegala, D. (2025). Investigating the Contextualised Word Embedding Dimensions Specified for Contextual and Temporal Semantic Changes. In *Proceedings of the 31st International Conference on*

Computational Linguistics (pp. 1413–1437). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2407.02820

[17] Worth, P. J. (2023). Word Embeddings and Semantic Spaces in Natural Language Processing. *International Journal of Intelligence Science*, 13(1), 1–21. https://doi.org/10.4236/ijis.2023.131001

[18] Palominos, C., He, R., Fröhlich, K., Mülfarth, R. R., Seuffert, S., Sommer, I. E., Homan, P., Kircher, T., Stein, F & Hinzen, W. (2024). Approximating the semantic space: word embedding techniques in psychiatric speech analysis. Schizophrenia, 10(1), 1-10.,

[19] Worth, P. J. (2023). Word Embeddings and Semantic Spaces in Natural Language Processing. *International Journal of Intelligence Science*, 13(1), 1–21. https://doi.org/10.4236/ijis.2023.131001

[20] Zhou, J., Liu, C., Duan, N., & Li, M. (2022). *An Overview of Pretrained Language Models for Natural Language Processing*. *AI Open*, 3, 9–28. https://doi.org/10.1016/j.aiopen.2021.12.001

[21] OpenAI. (2023). *GPT-4 Technical Report*. https://doi.org/10.48550/arXiv.2303.08774

[22] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2023). Language models are few-shot learners. *Communications of the ACM*, 66(5), 108–117. https://doi.org/10.1145/3571991

[23] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT 2019*. https://doi.org/10.48550/arXiv.1810.04805

[24] Wang, A., Zhang, Y., Liu, J., & Bowman, S. R. (2023). Evaluating Pretrained Transformers for Natural Language Understanding. *Transactions of the Association for Computational Linguistics*, 11, 245–261. https://doi.org/10.1162/tacl_a_00559

[25] Oancea, B. (2025). *Text classification using machine learning methods*. arXiv preprint arXiv:2502.19801. https://doi.org/10.48550/arXiv.2502.19801

[26] Abia, V. M., & Johnson, E. H. (2024). *Sentiment Analysis Techniques: A Comparative Study of Logistic Regression, Random Forest, and Naive Bayes on General English and Nigerian Texts*. Journal of Engineering Research and Reports, 26(9), 123–135. https://doi.org/10.9734/jerr/2024/v26i91268

[27] Shijaku, E., & Canhasi, E. (2024). *Classification of human- and AI-generated texts for different languages and domains*. International Journal of Speech Technology. https://doi.org/10.1007/s10772-024-10143-3

[28] Sanchez-Medina, J. J. (2024). *Sentiment analysis and random forest to classify LLM versus human source applied to Scientific Texts*. arXiv preprint arXiv:2404.08673. https://doi.org/10.48550/arXiv.2404.08673

[29] Makinde, H. S., Makinde, A. I., Usman, M. A., Adegoke, H., Makinde-Isola, B. A., Lawal, W., & Jimoh, I. T. The Readability Paradox: Can We Trust Decisions on AI Detectors? Technium Education and Humanities, 11, 181-195.

[30] Krawczyk, N., Probierz, B., & Kozak, J. (2024). Towards AI-Generated Essay Classification Using Numerical Text Representation. Applied Sciences, 14(21), 1-23.