European Journal of Education

WILEY

ORIGINAL ARTICLE OPEN ACCESS

A Comparison of Human-Written Versus AI-Generated Text in Discussions at Educational Settings: Investigating Features for ChatGPT, Gemini and BingAI

Hatice Yildiz Durak[1] | Figen Eğin[2] | Aytuğ Onan[2]

[1]Department of Educational Science, Eregli Faculty of Education, Necmettin Erbakan University, Konya, Turkey | [2]Department of Computer Engineering, Software Engineering, Izmir Katip Celebi University, İzmir, Turkey

Correspondence: Hatice Yildiz Durak (hatyil05@gmail.com)

Received: 18 October 2024 | **Revised:** 24 December 2024 | **Accepted:** 13 January 2025

Keywords: AI in education | AI-generated text | BingAI | chat generative pre-trained transformer (ChatGPT) | Gemini | human-written text | natural language processing

ABSTRACT

Generative artificial intelligence (GenAI) models, such as ChatGPT, Gemini, and BingAI, have become integral to educational sciences, bringing about significant transformations in the education system and the processes of knowledge production. These advancements have facilitated new methods of teaching, learning, and information dissemination. However, the widespread adoption of these technologies raises serious concerns about academic ethics, content authenticity, and the potential for misuse in academic settings. This study aims to evaluate the linguistic features and differences between AI-generated and human-generated articles in educational contexts. By analysing various linguistic attributes such as singular word usage, sentence lengths, and the presence of repetitive or stereotypical phrases, the study identifies key distinctions between the two types of content. The findings indicate that human-generated articles exhibit higher average singular word usage and longer sentence lengths compared to AI-generated articles, suggesting a more complex and nuanced language structure in human writing. Furthermore, the study employs ensemble learning models, including Random Forest, Gradient Boosting, AdaBoost, Bagging, and Extra Trees, to classify and distinguish between AI-generated and human-generated texts. Among these, the Extra Trees model achieved the highest classification accuracy of 93%, highlighting its effectiveness in identifying AI-generated content. Additionally, experiments using the BERTurk model, a transformer-based language model, demonstrated a classification accuracy of 95%, particularly in distinguishing human-generated articles from those produced by Gemini. The results of this study have significant implications for the future of education, as they underscore the critical need for robust tools and methodologies to differentiate between human and AI-generated content.

1 | Introduction

Generative artificial intelligence (AI) tools have become very popular technologies in everyday life (Mindner, Schlippe, and Schaaff 2023; Yildiz Durak and Onan, 2023). The massive increase in the development and deployment of large-scale Natural Language Generation systems in recent months has led to an unprecedented transformation (Herbold et al. 2023; Myers et al. 2024). Rapid developments in generative AI technologies have led to an increased interest in their capabilities and applications, especially in written work and transactions (Landa-Blanco, Flores, and Mercado 2023). These technologies have the ability to mimic human-like conversations with users, such as providing information and assistance, offering emotional support (Falala-Séchet et al. 2019; Yildiz Durak 2023).

The most frequently used of these technologies is OpenAI's ChatGPT. In addition to these technologies, other notable AI models include Gemini and BingAI (Eğin et al. 2025). Each of these technologies offers unique features and functionalities for generating human-like text. These technologies have been trained on large amounts of data to improve the ability to understand and generate natural language text and to understand and respond to natural language input (Ariyaratne et al. 2023). These technologies are diverse in terms of the users who use them for a variety of purposes, from researching educational content to everyday conversations and medical advice. But in particular, AI in education offers significant opportunities and innovations in several areas. The role of this technology in education ranges from improving teaching methods to personalising learning experiences. As these technologies become more prevalent in education, it is crucial to understand how they compare to human-generated content, especially in academic and educational contexts.

Therefore, as these systems have become part of our daily lives, it has become more important to distinguish between human- and AI-generated content. Human-generated content is created with the intention of reflecting intent. AI-generated content, on the other hand, is created by algorithms designed to generate text in response to prompts. AI-generated text may contain repetitive or stereotypical phrases or patterns, while human-generated text is more likely to be original and creative. Moreover, texts generated by large language models often appear to be reliable. Dergaa et al. (2023) and Van Dis et al. (2023) emphasise that technologies such as ChatGPT have the ability to generate well-written student essays, summarise research papers, answer questions, and generate useful computer code. There is also the potential to produce research summaries that are difficult to distinguish at first glance from those written by scientists, and these features are largely seen as concerning.

Given the diverse uses and potential of these technologies in education, there is a need for the research community to engage in a comprehensive discussion about the potential uses, threats and limitations of these tools. In addition to these needs, serious problems can arise in many areas, such as unethical situations and the production of fake content. Therefore, there is a need for tools that can distinguish between human and AI-generated texts. Therefore, in the current study, we focus on the differences between human-generated texts and AI-generated texts. Furthermore, this study aims to fill the existing research gap by analysing the effectiveness of generative AI technologies in writing argumentative texts and comparing them with student-generated texts.

1.1 | Research Questions

This paper addresses the current research gap regarding the comparison of generative AI technologies with human writing, guided by the following research questions.

RQ1: How do AI-generated articles compare to student-written articles?

RQ2: What are the linguistic tools that are characteristic of student- and AI-generated content?

RQ3: Can discussion-based articles generated by generative AI technologies based on ChatGPT, Gemini and BingAI be distinguished from human-generated articles?

1.2 | Related Work

There has been a recent increase in interest in generative AI models and these technologies are widely used in the form of interactive conversations (Ouyang et al. 2022; Yildiz Durak and Onan 2024). In particular, similar large generative language models based on transformer architecture, such as GPT-3, have not been trained in a self-monitoring manner (Brown et al. 2020; Herbold et al. 2023). Therefore, in order to reduce biases and ethical issues, to make the generated text more similar to the text written by humans, and to improve the quality of the outputs, comparisons with texts containing human-written content are necessary. When the studies conducted in this direction are examined, it is remarkably concluded that although generative AI technologies have a significant potential in academic writing as a tool, their current capabilities in terms of reference reliability, validity context, and so forth in academic text generation are limited. Studies in the literature have compared AI generated versus human generated content by examining different linguistic tools, language features, and human expert evaluations.

Cai et al. (2023) concluded that ChatGPT reuses sentence structure, accesses the intended meaning of an ambiguous word, and mimics human language use by identifying the thematic structure and arguments of a verb. Zhao et al. (2023) found that ChatGPT produced longer and more varied responses when the user was in a negative emotional state. The study by Mindner, Schlippe, and Schaaff (2023) addresses the challenge of identifying AI-generated texts in educational settings where tools such as ChatGPT are increasingly used to produce academic content. Several detection systems were developed and validated using a mix of traditional and novel features such as confusion, semantics, list searches, error patterns, readability metrics, AI feedback, and text vector analysis. These systems were tested on a newly created corpus covering 10 school subjects. The findings show that the detection systems achieved F1 scores of over 96% for distinguishing between human-generated and AI-generated texts, and over 78% for AI-rephrased texts. Remarkably, the performance of the base text rephrases detection system reached a value of 183.8% on the GPTZero model F1-score. It emphasised the effectiveness of integrated features in improving the accuracy of classifiers. This study shows that improved detection capabilities can play an important role in maintaining academic integrity in the face of advancing AI text generation technologies. Zhang and Zhang (2024) focus on extracting problem and method sentences to learn the main idea from academic papers. Steiss et al. (2024) compared AI and human feedback statements.

The study by Ariyaratne et al. (2023) compares the accuracy and quality of various academic articles generated by ChatGPT with those written by human authors. It evaluated the accuracy of ChatGPT-generated radiology articles by comparing them with articles that have been published or written and are under review. Two specialty-trained musculoskeletal radiologists performed the analysis and used a rating from 1 to 5. Overall, ChatGPT is able to produce consistent research articles that closely resemble actual articles published by academic researchers on initial review. However, all articles evaluated were factually incorrect and had fictitious references.

The study by Herbold et al. (2023) systematically evaluates the quality of AI-generated content through a large-scale study comparing human-generated and ChatGPT-generated discussion-oriented student essays. Discussions were evaluated by a large number of humans (teachers). The analysis is extended by taking into account a number of linguistic features of the generated texts. According to the results, ChatGPT produced discussion texts that scored higher in quality than human-written essays. The writing style of the AI models exhibited different linguistic features than the human-written texts. The aim of the study by Dergaa et al. (2023) was to investigate the potential benefits and threats of ChatGPT and other NLP technologies in academic writing and research literature, and to highlight ethical considerations related to the use of these tools. It also examines the impact that AI tools can have on the authenticity and credibility of academic work. This study conducts a literature review of scientific articles published in journals indexed as Q1 in Scopus using keywords such as "ChatGPT", "AI-generated text", "academic writing" and "natural language processing". Using a semi-qualitative approach, this study found that ChatGPT and other NLP technologies have the potential to improve academic writing and research productivity. However, the use of these technologies in academic writing raises concerns about the authenticity and reliability of the work.

Van Veen et al. (2023) examined the functionality of large language models (LLMs) in natural language processing (NLP) tasks, such as sifting very large textual data and summarising important information from electronic health records. The study applied adaptation methods to LLMs in eight domains covering six datasets and four different clinical summarization tasks. Ratings were provided by human experts. The results showed that the LLM outperformed the LLM in clinical text summarization across multiple tasks. Landa-Blanco, Flores, and Mercado (2023) analysed whether humans can rate creative writing texts differently if they believe the author is an AI (ChatGPT) or a person. An experimental design was structured. AI-generated texts were presented to a control and an experimental group. The content, including three poems and a short story, was generated by ChatGPT. Participants in the control group were told that the texts were written by a person, while the experimental group was told that the texts were generated by ChatGPT. There was no statistically significant difference between the scores of the control and experimental groups in terms of the perceived creativity and originality of the texts, how enjoyable they were, and the likelihood of the participants to recommend the texts to someone else. The conclusion was that, in terms of creative writing, readers did not rate a text attributed to human authorship differently from a text believed to have been written by an AI. Johansson (2023) aimed to compare an essay written by a student for an English literature course with an equivalent essay generated through ChatGPT. It was investigated whether the AI met the formal requirements of academic writing and its distinctiveness in the generated text in terms of "assertiveness, self-identification and authorial presence". The results showed that ChatGPT is able to produce seemingly appropriate context-based texts, but needs improvement in terms of actual accuracy and the subtle features of authorship found in human writing. AI-generated text lacks the depth, specificity and accurate source referencing present in human-generated text. This study concludes that while AI has potential as a tool, its current capabilities are limited, especially in academic text generation.

2 | Methods

This study aims to examine the linguistic features and differences between the articles of faculty of education students on the discussion topics determined on current issues in the field of instructional technologies and the articles produced by generative AI tools on the same tasks. For this purpose, a qualitative research method was used in the study. In qualitative research, a determined situation is analysed and explained in detail (Denzin and Lincoln 1998).

In this study, each of the discussion articles written by the students of the faculty of education was considered and analysed as a case. Thus, the case study method was used in the study. In a case study, the phenomenon is examined in a characteristic integrity with patterns and relationships. Therefore, in case studies, an in-depth explanation and analysis is made with a holistic approach about a selected event or situation (Merriam 1998). Within the scope of the research, articles written by students and generative AI tools in the context of 30 discussion topics were examined separately and a holistic examination was carried out on more than one case. Yin (2003) emphasised that case studies can be handled in different categories according to the purpose and number of cases. In this study, a holistic multiple case study was adopted (see Figure 1). In this design, each purposively selected case is handled holistically within itself and then compared with each other (Yıldırım and Şimşek 2013; Yin 2009).

2.1 | Participants

The participants consisted of 128 students from the faculty of education. The average age of the participants is

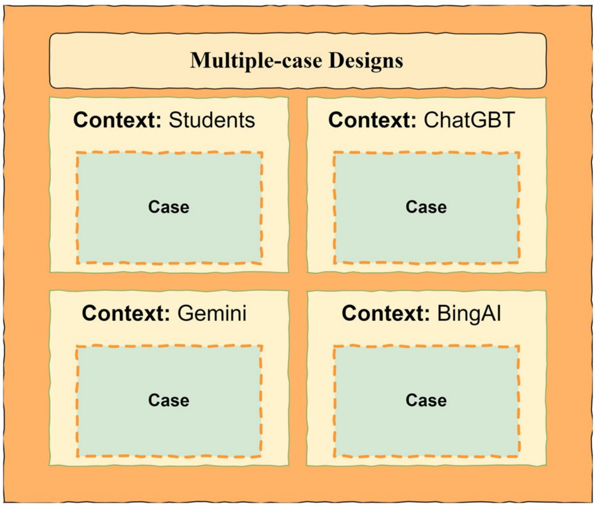

FIGURE 1 | The research design.

approximately 20 years old. All participants are enrolled in a 14-week Instructional Technology course. The students were given an overview of language models and basic information about how Gemini, BingAI and ChatGPT generate content. They were also given both written and verbal explanations on how to write articles for the purpose of this study. Students were given 30 days to complete the study tasks. Participation in the study was voluntary and all data were collected anonymously. Approximately 15% of the students enrolled in the course did not participate in the study.

2.2 | Data Collection Form

The sections written by the students in the context of the given discussion topics were collected with an online form. This form consists of two parts. In the first part, demographic information was collected. In the second part, 30 discussion topics and the answers generated by ChatGPT 3.5, Gemini and BingAI in the context of these topics were collected. ChatGPT, which has strong language skills, is used in different applications in education (Yeşilçınar 2023). Gemini AI, created by Google DeepMind, draws attention with its versatility in different data methods (Perera and Lankathilake 2023); Bing AI has creative, sensitive and balanced speech options (Salazar et al. 2023). The reason for choosing these three models in this study is that, in addition to these features, they are frequently used in education. ChatGPT and Gemini are the most widely used AI tools (Bayer, Ince Araci, and Gurkan 2024). ChatGPT reached one million users in just 5 days, and Bing AI is one step ahead of ChatGPT in accessing up-to-date data thanks to internet access (Motlagh et al. 2023). The questionnaire used in the study was structured as shown below and sample responses are presented in Table 1.

2.3 | Datasets

The dataset consists of human-generated texts from students' answers to 30 discussion topics related to technology use and AI-generated texts from the same questions entered as prompts in ChatGPT, Gemini, and Bing AI. The number of participants in the dataset is high to ensure diversity in writing styles in human-written texts. In AI-written texts, three separate language models (ChatGPT, Gemini and Bing AI) provide this diversity.

The dataset comprises 121 rows and 121 columns. The columns contain 30 discussion topics for humans and three AI models. The rows contain the title row, the answers from the participants, and the responses they received from the prompts they entered into the three AI tools.

2.4 | Analysis and Models

In order to compare the argumentative articles created using GenAI technologies based on ChatGPT, Gemini and BingAI with the articles written by students, text preprocessing steps were performed first. Text preprocessing includes cleaning the text from punctuation marks, numbers, and stopwords. After preprocessing, the features of the text are extracted. In this step, separate number of sentences, number of conjunctions, unique word count, number of stopwords are extracted for human-written and AI-written texts. In addition, the number of sentences, unique words, stopwords, sentence lengths (word & characters), and word lengths (characters) values are extracted. N-grams are created by tokenizing the words in the text, dividing them into groups of n numbers, and classifying them in this way (Cavnar and Trenkle 1994). In other words, bigrams represent word pairs in the text, and unigrams represent individual words. This study calculated the frequencies of unigrams and bigrams in the text, and the first 15 with the highest frequency were listed. All these steps were performed using the Python NLTK library. In addition, word clouds were created with WordCloud and matplotlib libraries.

In the next step, TF-IDF was applied to vectorize the text. TF-IDF is a prevalent method for determining the weights of words in the text (Yun-tao, Ling, and Yong-cheng 2005; Yildiz Durak and Aydoğdu, 2023). Another Python library, Sklearn, was used to implement TF-IDF. The TF-IDF approach was used for feature extraction and Random Forest, Gradient Boosting, AdaBoost, Bagging and Extra Trees ensemble learning models were used with their standard classifiers without changing any parameters (see Figure 2). The ensemble learning models selected for this study, including Random Forest, Gradient Boosting, AdaBoost, Bagging, and Extra Trees, were chosen for their complementary strengths and proven effectiveness in handling high-dimensional and complex datasets. These models leverage multiple learners to improve performance and exhibit diverse features that make them well-suited for the task. Random Forest, known for its robustness, performs exceptionally well on high-dimensional datasets and effectively mitigates overfitting by combining multiple decision trees (Schoppa, Disse, and Bachmair 2020). Extra Trees Classifier provides computational efficiency and reduces overfitting through random feature selection, offering faster training times (Darbanian, Rahbari, Ghanizadeh & Nickray, 2020). Gradient Boosting, with its ability to sequentially build on weak learners, prevents overfitting and offers flexibility for a variety of scenarios (Bentéjac, Csörgő, and Martínez-Muñoz 2021). AdaBoost, grounded in a strong theoretical foundation, improves prediction accuracy by focusing on misclassified samples, particularly in classification problems (Ying et al. 2013; Darıcı 2023). Bagging further enhances classification performance by reducing variance and increasing model stability through multiple dataset resamplings (Sandag 2020). Together, these features played a critical role in the selection of these models.

For this study, the models were initially run with default hyperparameters to establish baseline performance and provide a clear understanding of their intrinsic strengths. This straightforward approach ensured meaningful comparisons without introducing additional complexity. Future work will focus on exploring advanced hyperparameter tuning techniques, such as grid search or Bayesian optimization, to further enhance model performance and robustness. While extracting the features, 1000 words with the highest TF-IDF score were selected and vectorization was performed over these words. Using the "train_test_split" function in the "scikit-learn" library, 20% of the dataset was split as test dataset and 80% was used as training set.

In the second stage, the articles were cleaned from stopwords, special characters and numbers and binary classifications were made between human and AI produced articles with BERTurk, one of the transformers-based models (see Figure 3). For the

TABLE 1 | Example of instruments and responses.

Model	Context	Prompt example	Responses
Human	How would you relate the concepts of Education 4.0, Industry 4.0 and Society 5.0?	—	They are all about development, one is about education, one is about mechanisation, one is about social sciences, but the main purpose of all of them is to improve the relevant field. To increase the level of welfare. In the Industry 4.0 process, digitalization, artificial intelligence, automation, etc. We must master technology because technology is everywhere in our lives at the present time. From public transportation systems to cooking, we are intertwined with technology even in everyday tasks that we can think of. Being technologically literate makes even our daily lives easier and adapts to the current world. In this way, we create individuals who are more creative, more productive and able to manage information. In this way, they will be more conscious while protecting their own data, and they will plan accordingly and acquire activities for protection
ChatGPT3.5	What is the role of Industry 4.0 and digital transformation in education?	"You are a student at the faculty of education. How would you relate the concepts of Education 4.0, Industry 4.0 and Society 5.0? Write an opinion piece of at least 100 words"	Industry 4.0 and digital transformation bring about a technology-driven change in education. This transformation adapts to the needs of the labour market by equipping students with digital skills. The use of digital tools and artificial intelligence in education personalises learning experiences and improves students' problem-solving abilities. It also enables new learning methods such as distance learning and online resources. In this way, it contributes to the development of students as competitive and continuous learners who are compatible with technology. Industry 4.0 and digital transformation also change the role of teachers in education, transforming them into guides who can use technology effectively
Gemini	In which areas can artificial intelligence be used in education?		Artificial intelligence (AI) can be used in a wide variety of areas in education. Some of these include: • *Personalised learning*: AI can be used to deliver learning experiences tailored to each student's individual needs and learning styles • *Automated assessment*: AI can be used to automatically assess assignments, tests and other assessments. This allows teachers to save time and give feedback to students faster • *Student support*: AI can be used to help students with their assignments, answer their questions and increase their motivation • *Data analysis*: AI can be used to analyse educational data and learn about student performance. This information can be used to improve educational programs and teaching practices • *Summary*: AI has the potential to make education more effective and efficient. The use of AI in education is still in its infancy, but it has the potential to transform education in the future
BingAI	What are the main differences between Web 1.0, Web 2.0 and Web 3.0?		The main differences between Web 1.0, Web 2.0 and Web 3.0 are as follows: Web 1.0: The first iteration is a "read-only web". Users view information by forcing it through servers. The relationship is one-sided. Web 2.0: Information passes in multiple direction. Users can interact and create content. It is the beginning of a collaborative environment. Web 3.0: The goal is to decantralize data and reclaim users' digital identities.

implementation of the BERTurk model, the Numpy library was used, and the datasets were randomly divided into 80% training and 20% testing, and the training was completed in 3 epochs. In the implementation of the model, the learning rate parameter (lr) was chosen as "3e-5".

2.5 | Random Forest

The Random forest model, one of the most successful classifiers in machine learning models, is a machine learning algorithm based on a collection of decision trees. This model provides

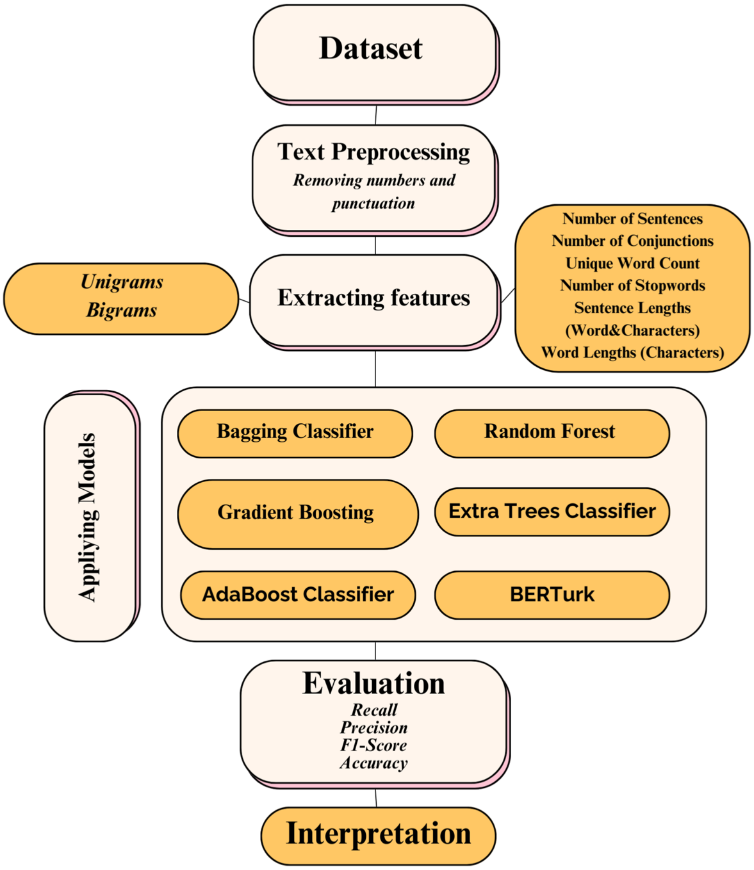

FIGURE 2 | Analysis procedure.

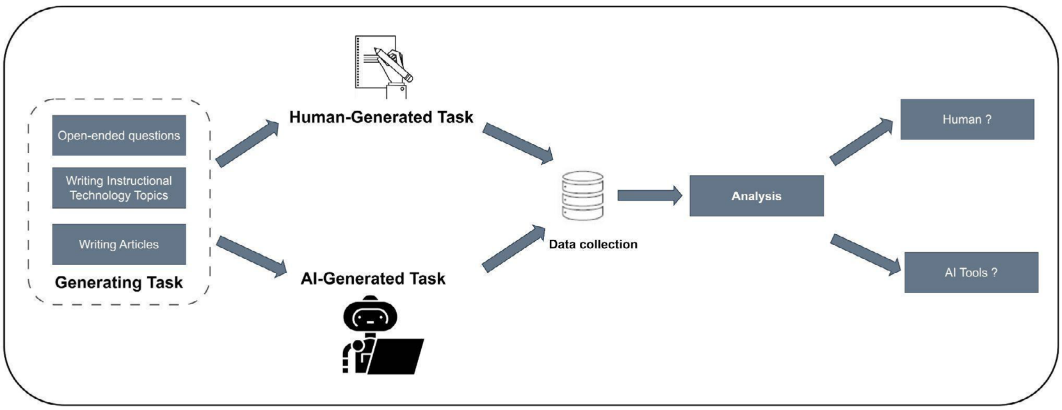

FIGURE 3 | Research process.

high accuracy in various tasks such as classification and regression by combining many decision trees, each trained using a different data subset and feature set (Durak and Bulut 2024; Breiman 2001).

2.6 | AdaBoost Classifier

Boosting, an important method in ensemble learning, combines weak learners into strong learners. Its ability to be applied to almost all machine learning models has paved the way for its frequent use to improve prediction accuracy. This model, called Adaptive Boosting because it can be adapted to the training error of each base classifier, is primarily designed for the classification problem (Ying et al. 2013).

2.7 | Gradient Boosting Classifier

Gradient Boosting is an ensemble learning model based on the negative gradient of the loss function when adding new models. The task starts with a simple model, and in each cycle a new model is added to correct the model's errors. In this model, where task-specific loss functions can be applied, errors are corrected successively during the learning process for more accurate prediction (Natekin and Knoll 2013).

2.8 | Bagging Classifier

Bagging is based on training the same machine learning algorithm on randomly generated subsets of the current dataset and using these models to obtain the final prediction through voting or averaging. The generated subsets of the dataset are called bags (Ngo, Beard, and Chandra 2022).

2.9 | Extra-Trees Classifier

Extra-Trees, an algorithm similar to the random forest algorithm, fits random decision trees to subsamples of the dataset to improve prediction accuracy. To prevent over-fitting, it uses mean value and meta-estimator (Darbanian, Rahbari, Ghanizadeh and Nickray; 2020).

2.10 | BERTurk

Language models using the Transformer architecture introduced by Vaswani et al. (2017) have attracted attention with their high performance. One of these pre-trained models, BERTurk, which uses the BERT infrastructure, was trained using a large Turkish corpus (Schweter 2020). With these features, it can be considered as a very useful model for Turkish language processing tasks.

3 | Results

This section presents the study's results and the research questions. How AI articles can be compared with student articles (RQ1) and what the characteristic features of the texts are (RQ2) are revealed by examining the syntactic and semantic features of the texts. Whether AI and human-written articles can be distinguished (RQ3) was answered with ensemble models and experimental studies conducted with BERTurk.

3.1 | Syntactic Features of the Human and AI-Writing Articles

In-depth analyses were conducted with the Python NLTK library to compare AI-written texts with human-written texts and to reveal the characteristic features of the texts. The result of these analyses, some attributes of the dataset such as total characters, words, sentences and conjunctions are presented in Table 2. It can be seen that the total number of sentences and conjunctions is lowest in student articles and highest in articles created by Gemini. The total number of unique words was the lowest in the articles created by the ChatGPT and the highest in the human-generated articles. Stopwords analysis showed that the highest use of stop word was in Gemini and the least use of stop word was in human-written articles.

Some text-based features were also examined, as it was observed that the human-generated sections were shorter in the dataset. For each input, the average number of sentences, the number of unique words, and the average length of sentences and words were determined and evaluated. The results of this evaluation are given in Table 3.

TABLE 2 | Features of the dataset.

Features	Human	ChatGPT	Gemini	BingAI
Number of sentences	21,718	44,815	61,033	47,269
Number of conjunctions	24,669	49,442	66,187	41,288
Unique word count	24,301	15,639	18,952	16,237
Number of stopwords	44,199	80,097	121,217	75,499

TABLE 3 | The mean values for features.

Features	Human	ChatGPT	Gemini	BingAI
Number of sentences	5.36	11.11	15.14	11.73
Unique word	65	21	19	23
Using stopwords	118	109	123	109
Sentence lengths (word)	17.271	16.376	16.119	14.633
Word lengths (characters)	6.413	6.341	6.102	6.051
Sentence lengths (characters)	110.757	103.833	98.350	88.548

The average values show that the average number of sentences per article is the highest in Gemini and the lowest in human-generated articles. The average number of unique words is very close in the AI-generated articles, but lowest in the Gemini articles. Human-generated articles have a significantly higher number of unique words (about 4 times). In terms of average stopwords usage, Gemini ranked first, followed by human-generated articles. When sentence lengths are analysed, human-generated articles rank first. The shortest sentences were created by BingAI. Although there is not much difference, the words created by BingAI also have fewer characters on average compared to the others. When the sentence lengths are analysed on the basis of characters, the order is human, ChatGPT, Gemini and BingAI from largest to smallest.

The analysis of the number of sentences in the AI and human-generated articles show that the human-generated articles are shorter, while the total number of unique words is higher in human articles (see Figure 4).

In the analysis of the average values, it was also found that the number of singular words was significantly higher in human articles. When the average number of sentences was analysed, it was found that AI articles used 2 times more sentences than human articles. While sentence lengths are higher in human articles in terms of both words and characters used, there is not much difference between word lengths. When compared to other articles, BingAI-generated articles have shorter sentences,

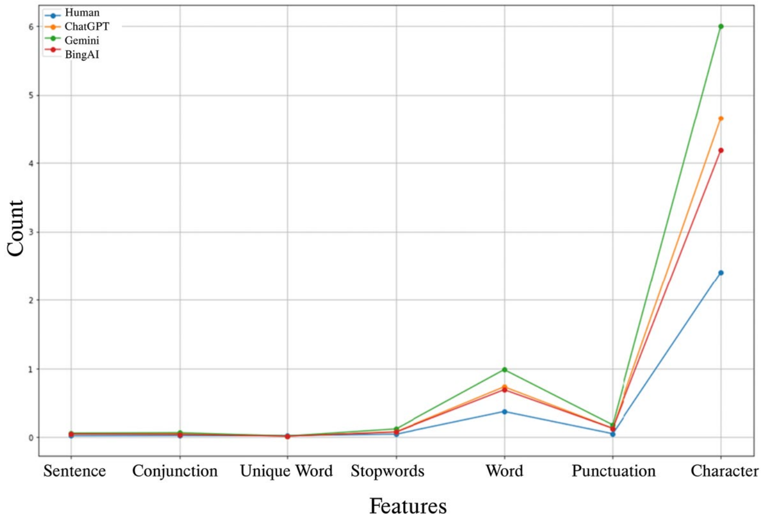

FIGURE 4 | Comparison of total values of features.

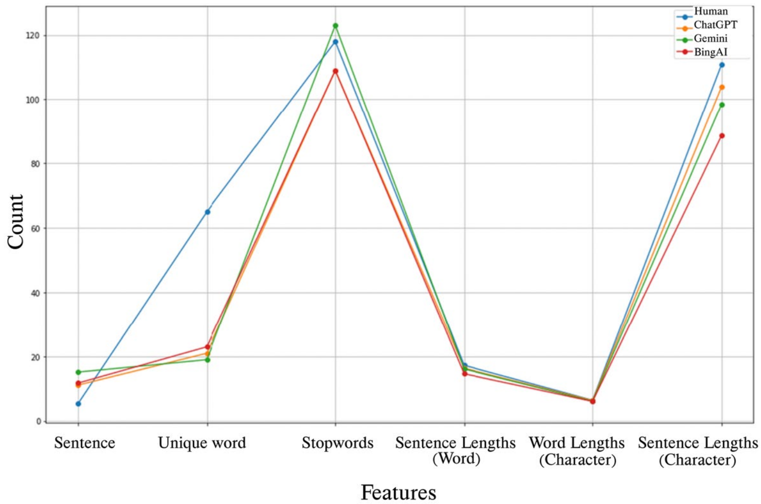

FIGURE 5 | Comparison average values of features.

and Gemini articles have more words and punctuation marks than others (see Figure 5). These findings, summarised in Figures 3–5, show that the distinction between human and AI can be evaluated by considering the syntactic values for RQ1 and RQ2. The use of stopwords in total values and the use of singular words in average values are an important distinguishing feature in terms of syntax. Syntactic complexity is assessed by the average sentence length and the number of characters per word (Beers and Nagy 2009). When evaluated from this perspective, human-written texts have higher syntactic complexity.

3.2 | Semantic Features of the Human and AI-Writing Articles

N-grams and word clouds reveal semantic differences between AI and human-written texts. In order to evaluate the dataset in semantic perspective, n-gram was also used. The frequencies of unigrams and bigrams of human and AI-generated articles were extracted separately and the 15 most frequent n-grams in the text were identified (see Figure 6). According to these results, human-written texts include more diverse topics such as "visual design" and "instructional technology," but AI texts show more repetition and standard expressions. Furthermore, some bigrams are frequently repeated in AI-generated texts (chatGPT, Bing and Gemini), indicating that specific patterns are more prevalent in AI-written texts.

The word clouds of the datasets allow the evaluation of prominent word groups. In this context, the variety of words used in human writing shows a more balanced distribution. Technical terms such as "artificial intelligence" and "operational thinking" and expressions such as "one way" and "visual design" are frequently used; human writings approach the subject from different perspectives and offer a variety of expressions (Figure 7). In the texts produced by ChatGPT, the expression "one way" is large and prominent. In addition, "artificial intelligence" and "operational thinking" are frequently used, but the other words are not prominent. This situation reveals that ChatGPT tends to repeat certain expressions and has a limited interpretation of word diversity (Figure 8).

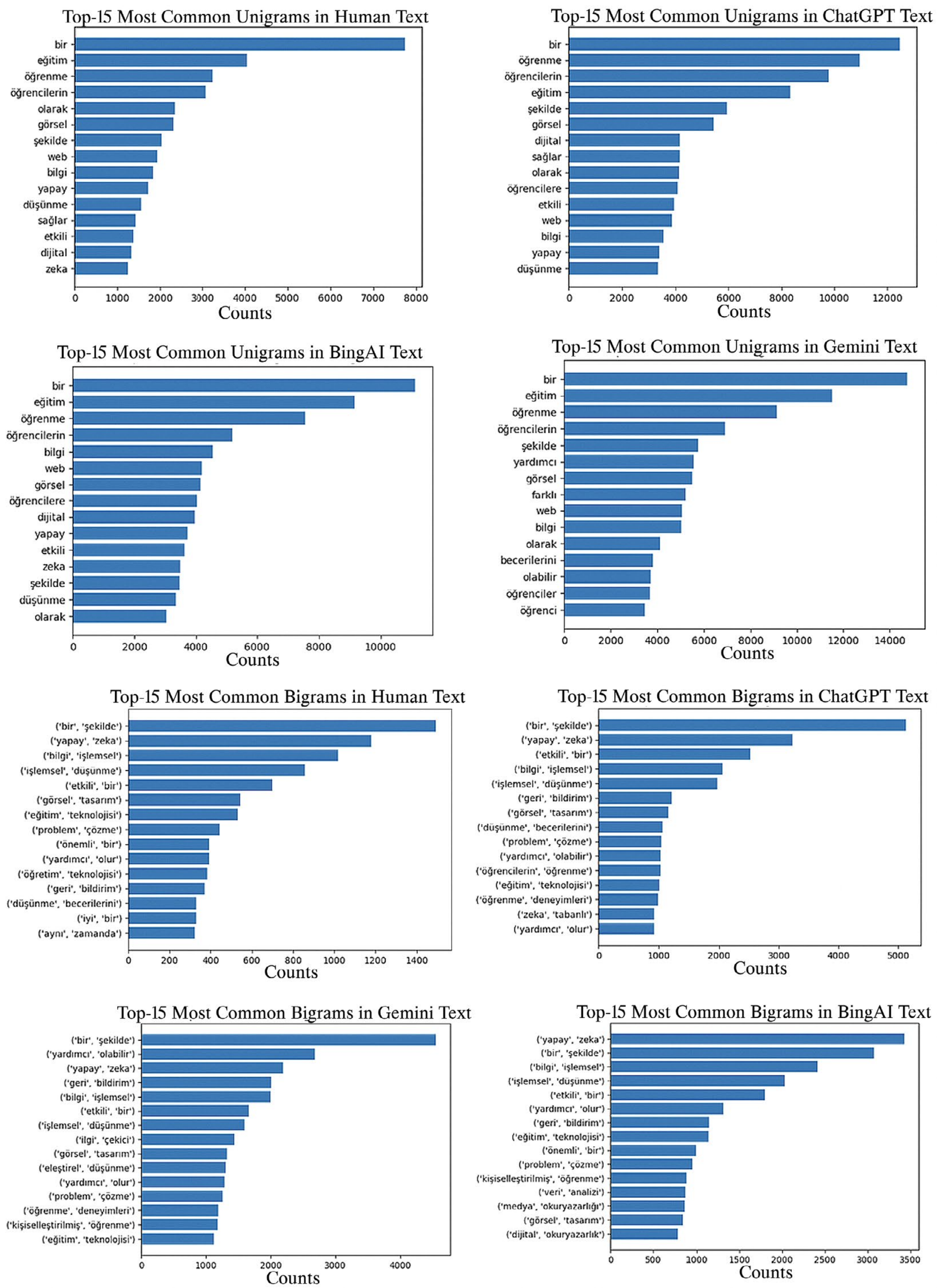

FIGURE 6 | Bigrams and unigrams of human and AI-generated text.

In the texts written by Gemini, the words "helpful," "can," and "one way" are prominent. Words like "artificial intelligence" and "feedback" appear more significant than human writings, suggesting that Gemini adopts a more user-focused language, adopting a "helpful" tone. Furthermore, the appearance of expressions such as "critical thinking" suggests that Gemini also emphasises different conceptual elements in its narrative (Figure 9). The expressions "artificial intelligence," "information operational," and "one way" seem to dominate the texts produced by BingAI. This distribution suggests that BingAI focuses on technical terminology but uses elements similar to human writing, such as "operational thinking" and "education technology" (Figure 10). Figures 6–10, when evaluated in terms of RQ3, show that human-AI written texts can be distinguished by taking into account repetitive content.

3.3 | Ensemble Models

In the experiments conducted with ensemble learning models to distinguish student articles from AI-generated articles,

FIGURE 7 | Human-writing text word cloud.

FIGURE 8 | ChatGPT-writing text word cloud.

FIGURE 9 | Gemini-writing text word cloud.

FIGURE 10 | BingAI-writing text word cloud.

TABLE 4 | Accuracy scores of ensemble models.

Model	Human-ChatGPT	Human-Gemini	Human-Bing
Random Forest	0.87	0.95	0.92
Gradient Boosting	0.85	0.93	0.90
AdaBoost	0.83	0.93	0.86
Bagging	0.85	0.93	0.88
Extra Trees	0.89	0.95	0.93

it was observed that the best accuracy scores were achieved with the Extra Trees model (see Table 4). The Random Forest model also achieved results as good as Extra Trees in distinguishing Gemini-generated articles from human articles. The lowest accuracy score was achieved with the AdaBoost model. The accuracy values obtained in detecting articles written with ChatGPT from human articles are lower than Gemini and BingAI. Gemini articles were the articles that could be detected with the highest accuracy score in each ensemble learning model (see Figure 11). The highest Precision, Recall, and F1 score values were achieved with the Human-Gemini data set. The highest Precision value was 0.97 with Random Forest, the highest Recall value was 0.95 with Extra Trees, and the highest F1 Score was 0.95 with Random Forest and Extra Trees models (see Tables 5–7).

The precision scores in Table 5 show the ability of the models to correctly classify a source as "Human" or "AI." When comparing the texts written by Human and Gemini, the highest performance was achieved mainly with Random Forest (0.97) and Extra Trees (0.96). Refers to the texts produced by Gemini, which can be distinguished more easily from human texts. In the experiments conducted with the texts written by Human and ChatGPT, the Gradient Boosting model (0.82) showed a low precision, which may mean that the texts written by ChatGPT are more similar to human texts and are more difficult to parse.

The Extra Trees model achieved the highest recall scores for the texts written by "Human and Gemini" (0.95) and "Human and ChatGPT" (0.93) (see Table 6), indicating that this model is more effective in distinguishing Gemini and ChatGPT texts from human texts. In the experiments conducted with human-written and Bing-written texts, the Bagging model, in particular, has a lower recall score (0.86), indicating that Bing AI texts are more difficult to distinguish from human texts. Bing texts need to be more distinguishable.

The F1 scores show that all models perform best in parsing "Human-Gemini" texts regarding precision and recall balance, suggesting that Gemini texts have a more distinct writing style and are easier to distinguish from human texts (see Table 7). In the experiments conducted with human-written and ChatGPT-written texts, the high F1 scores of the Random Forest (0.89) and Extra Trees (0.89) models, in particular, indicate that some models show consistent success in parsing ChatGPT texts. However, this success is more limited compared to other models.

3.4 | BERTurk Model

In the binary classifications made with the BERTurk model, the highest accuracy score was 0.95% in the classification of student-generated articles and Gemini-generated articles. The lowest classification accuracy was achieved with BERTurk in detecting human-generated articles from ChatGPT-generated articles with a success rate of 0.88% (see Table 8).

In the classification made between Human-ChatGPT, the 1st epoch accuracy score was 0.85%. The 2nd and 3rd epoch accuracy scores were found to be 0.86% and 0.88%, respectively. BERTurk accuracy score was 0.87%. For human-labelled data, Precision was calculated as 0.95%, Recall was calculated as 0.80%, and F1-score was calculated as 0.87%. For data labelled ChatGPT, Precision was found to be 0.82%, Recall was 0.96% and F1-score was 0.89%.

When Human-Gemini labelled data were classified with BERTurk, the 1st epoch accuracy score was 0.94%, the 2nd and

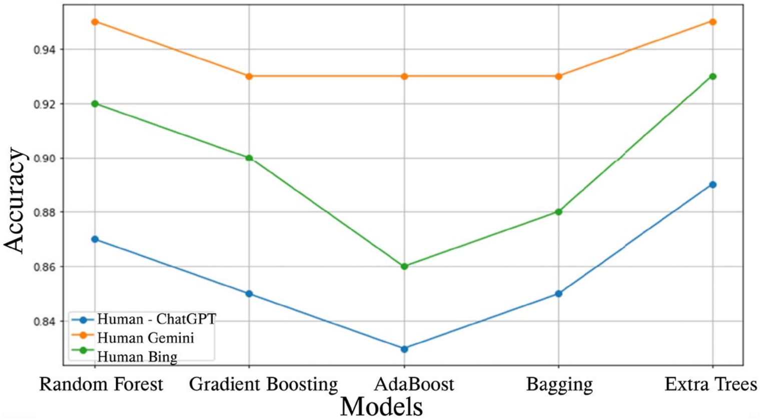

FIGURE 11 | Accuracy comparison of ensemble models.

TABLE 5 | Precision scores of ensemble models.

Model	Human-ChatGPT	Human-Gemini	Human-Bing
Random Forest	0.86	0.97	0.93
Gradient Boosting	0.82	0.94	0.91
AdaBoost	0.89	0.93	0.86
Bagging	0.85	0.94	0.89
Extra Trees	0.85	0.96	0.94

TABLE 6 | Recall scores of ensemble models.

Model	Human-ChatGPT	Human-Gemini	Human-Bing
Random Forest	0.91	0.93	0.90
Gradient Boosting	0.89	0.92	0.88
AdaBoost	0.85	0.93	0.85
Bagging	0.84	0.90	0.86
Extra Trees	0.93	0.95	0.91

3rd epoch was 0.96%. BERTurk classification accuracy is calculated as 0.95%. For human-labelled data, Precision was calculated as 0.94%, Recall was 0.96% and F1-score was 0.95%. For Gemini-labelled data, Precision was found to be 0.96%, Recall was 0.94% and F1-score was 0.95%.

In the Human-Bing dataset, the last classification task carried out with BERTurk, the 1st epoch accuracy score was 0.91, and the value increased to 0.92% in the 2nd epoch. In the 3rd epoch it was 0.94%. BERTurk accuracy score is calculated as 90%. For human-tagged data, Precision was 0.95%, Recall was 0.86%, and F1-score was 0.90%, while for Bing-tagged data, Precision was calculated as 0.87%, Recall was 0.95%, and F1-score was 0.91%.

Precision, Recall and F1-Score and accuracy scores of the experimental studies conducted with BERTurk are summarised in Figure 12 and Figure 13.

TABLE 7 | F1 scores of ensemble models.

Model	Human-ChatGPT	Human-Gemini	Human-Bing
Random Forest	0.89	0.95	0.92
Gradient Boosting	0.85	0.93	0.89
AdaBoost	0.83	0.93	0.86
Bagging	0.85	0.92	0.88
Extra Trees	0.89	0.95	0.93

4 | Discussion

In this study, findings indicate significant linguistic differences between human-generated and AI-generated content. The count of unique words is notably higher in human articles. However, there is no significant difference in word lengths between AI-written and human-written discussions. Articles generated by BingAI, when compared with others, feature shorter sentences, while Gemini articles have more words and punctuation than the rest. Additionally, a study by Herbold et al. (2023) has determined that AI-generated articles are highly structured linguistically, frequently using similar patterns. ChatGPT models have been found to be stricter in achieving this by creating clear logical structures through paragraphing different arguments instead of using discourse markers, thereby reducing the need for such markers. Ensemble learning models have been used to distinguish student articles from those created by AI, with the Extra Trees model achieving the best accuracy. In distinguishing Gemini-generated articles from human articles, the Random Forest model has performed as well as the Extra Trees model. Articles written with ChatGPT 3.5 have shown lower accuracy in distinction from human articles compared to those by Gemini and BingAI. Gemini articles have been consistently distinguished with the highest accuracy across all ensemble learning models, indicating the robustness and effectiveness of these models in text classification. Conversely, the lower classification accuracy of ChatGPT 3.5 texts may suggest that this model's language generation capability is more similar to human writing. According to studies by Landa-Blanco, Flores,

TABLE 8 | Results of BERTurk model (human = 0, AI models = 1).

Dataset	Label	Precision	Recall	F1-Score	Accuracy
Human-ChatGPT	0	0.95	0.80	0.87	0.88
	1	0.82	0.96	0.89	
Human-Gemini	0	0.94	0.96	0.95	0.95
	1	0.96	0.94	0.95	
Human-BingAI	0	0.95	0.86	0.90	0.91
	1	0.87	0.95	0.92	

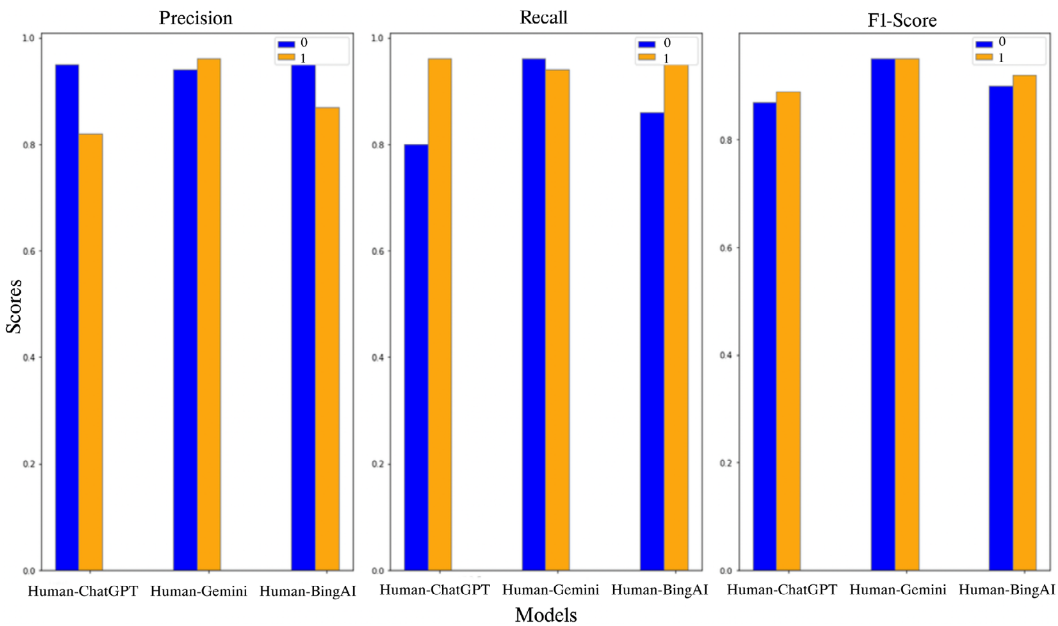

FIGURE 12 | Human-AI comparison according to evaluation metrics.

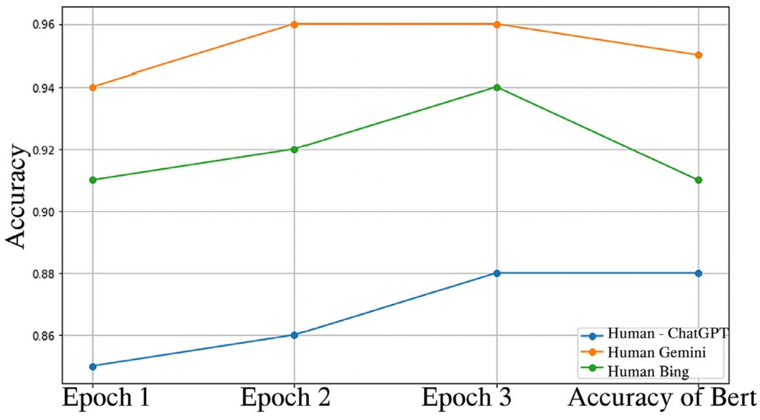

FIGURE 13 | Accuracy comparison of BERTurk.

and Mercado (2023) and Hitsuwari et al. (2023), people have assessed AI-generated creative writings as aesthetically similar to human writing. In binary classifications conducted with the BERTurk model, the highest performance, with an accuracy rate of 0.95, was achieved in classifying student writings from Gemini-generated articles. The lowest performance was in distinguishing human-written articles from those written by ChatGPT 3.5, with an accuracy of 0.88 using BERTurk. Particularly, BERTurk achieved a very high accuracy rate of 95% between Gemini articles and student writings, demonstrating the potential strength of language-based deep learning models in such tasks. These findings underscore the critical importance of model selection in the automatic detection and evaluation processes of AI-generated content.

AI tools have become quite widespread in education and are used to increase the quality of education in different fields (Chen, Chen, and Lin 2020). AI can be an important auxiliary tool when evaluated from the perspective of both students and teachers. Teachers experience low teaching quality and job commitment due to emotional exhaustion (Wang, Xin, and Chen 2024; Zhao and Wang 2024), and receiving support from AI tools for tasks such as evaluation can be an important support that can play a role in increasing the quality of education. The use of AI tools in education to support the development of students is inevitable. Today, students use AI tools to improve their writing skills, eliminate grammatical errors in their articles, and increase comprehensibility and consistency thanks to real-time feedback (Wu, Wang, and Wang 2024). Moreover, AI tools also positively affect students' skill development. A study conducted by Liu and Wang (2024) showed the effects of AI tools in developing critical thinking skills. According to the results of this study, personalised feedback and interactive learning environments provided by AI deepen students' writing and discussion skills. Despite its potential benefits, students using AI to perform tasks for their academic development may negatively affect their education.

In this context, the ethical use of AI and the separation of AI-generated texts from human-written texts are important. Tools such as Turnitin need to be sufficient to identify AI-generated texts. A study revealed that only 54.8% of AI texts could be correctly identified (Perkins et al. 2024). Encouraging students to use AI tools to support the learning process is important. Students should see these tools as an aid for skill development. The effective use of AI in education seems possible with students' ethical concerns and the development of practical tools that detect AI. With AI detection tools, it is possible to understand to what extent students use these tools. Thus, it is possible for students to benefit from these tools in a controlled manner. These research results show that ensemble models such as Extra Trees and Random Forest or the BERTurk model are successful in AI detection. AI detection tools using these models can distinguish human-written texts from AI-written texts with high accuracy. Software based on these models in particular can be produced for AI detection for educational institutions.

Additionally, instructors use these tools to evaluate student work (Chen, Chen, and Lin 2020). This situation carries some risks, such as data security. Furthermore, AI tools such as ChatGPT can make biased evaluations based on educational corpora. This situation can put students whose native language is not English at a disadvantage.

Studies show that classroom environments become more effective when teachers establish positive relationships with students (Wang and Wang 2024). Therefore, since this is not something that can be achieved with AI tools, AI tools should always be in a supporting position and should be used to support in every stage of the learning process.

5 | Conclusion, Limitations and Future Study

In conclusion, this study investigates new features for detecting articles written by humans versus those generated by AI tools. A new dataset has been created for this research. Different ensemble models and the BERTurk model were employed in the study. High accuracy rates have been achieved for both human-written and AI-generated base texts. ChatGPT has been found to produce content most similar to human writing. With tools like ChatGPT being readily accessible today, the assessment of exams or student assignments poses a significant challenge in terms of measurement and evaluation. Indeed, this situation could harm students' learning outcomes and academic integrity. According to the performance of the models on texts generated by different AI tools, Gemini-generated texts were best distinguished from human-written texts. All models performed better on the Gemini-Human dataset, suggesting that ChatGPT and Bing AI-generated texts are closer to human style than Gemini.

AI-written texts differ from human-written texts in terms of syntactic use of sentences, singular words, conjugations, and stop words. Sentence length and word length are similar. The findings could significantly contribute to the detection of AI-generated texts and also assist teachers in identifying generated content. Furthermore, our findings are not limited to academic writing but can also be applied to other areas such as creative writing and technical documentation.

This study has several limitations. To better understand the differences in argumentation between human-written and AI-generated article content, a more systematic study on semantic structures is needed. Additionally, assessments of the quality of student-written articles could be beneficial. The results were obtained based on undergraduate student papers. Students with a graduate-level education or expertise in specific fields are likely to perform better. In such cases, the performance gap between AI models and humans may differ. Moreover, different results may be obtained from participants from different ages, education levels, and cultures. Therefore, it would be appropriate to conduct similar studies with different participant groups in the future. It cannot be conclusively determined that the results can be generalised to similar models like ChatGPT-3.5, Gemini, and BingAI. Generalising and predicting outcomes, especially linguistic features, is challenging. Finally, it is recommended to bear in mind that the evolving structure and rapid progress of GenAI represent moving targets.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

Ariyaratne, S., K. P. Iyengar, N. Nischal, N. Chitti Babu, and R. Botchu. 2023. "A Comparison of ChatGPT-Generated Articles With Human-Written Articles." *Skeletal Radiology* 52, no. 9: 1755–1758.

Bayer, H., F. G. Ince Araci, and G. Gurkan. 2024. "ChatGPT-4o, ChatGPT-4 and Google Gemini Are Compared With Students: A Study in Higher Education." *International Journal of Technology in Education and Science* 8, no. 4: 627–644.

Beers, S. F., and W. E. Nagy. 2009. "Syntactic Complexity as a Predictor of Adolescent Writing Quality: Which Measures? Which Genre?" *Reading and Writing* 22: 185–200.

Bentéjac, C., A. Csörgő, and G. Martínez-Muñoz. 2021. "A Comparative Analysis of Gradient Boosting Algorithms." *Artificial Intelligence Review* 54: 1937–1967.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45: 5–32.

Brown, T., B. Mann, N. Ryder, et al. 2020. "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33: 1877–1901.

Cai, Z. G., D. A. Haslett, X. Duan, S. Wang, and M. J. Pickering. 2023. "Does ChatGPT Resemble Humans in Language Use?" arXiv preprint arXiv:2303.08014.

Cavnar, W. B., and J. M. Trenkle. 1994. "N-Gram-Based Text Categorization." In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval (161175, 14. USA).

Chen, L., P. Chen, and Z. Lin. 2020. "Artificial Intelligence in Education: A Review." *IEEE Access* 8: 75264–75278.

Darıcı, M. B. 2023. "Performance Analysis of Combination of Cnn-Based Models With Adaboost Algorithm to Diagnose Covid-19 Disease." *Politeknik Dergisi* 26, no. 1: 179–190.

Darbanian, E., D. Rahbari, R. Ghanizadeh, and M. Nickray. 2020. "Improving Response Time of Task Offloading by Random Forest, Extra-Trees and Adaboost Classifiers in Mobile Fog Computing." *Jordanian Journal of Computers and Information Technology* 6, no. 4: 345–360.

Denzin, N. K., and Y. S. Lincoln. 1998. *Strategies of Qualitative Inquiry*. Thousand Oaks: Sage.

Dergaa, I., K. Chamari, P. Zmijewski, and H. B. Saad. 2023. "From Human Writing to Artificial Intelligence Generated Text: Examining the Prospects and Potential Threats of ChatGPT in Academic Writing." *Biology of Sport* 40, no. 2: 615–622.

Durak, A., and V. Bulut. 2024. "Classification and Prediction-Based Machine Learning Algorithms to Predict Students' Low and High Programming Performance." *Computer Applications in Engineering Education* 32, no. 1: e22679.

Eğin, F., A. Onan, and H. Yildiz Durak. 2025. "Let's Talk About EdTech! A Topic Modelling Analysis of AI Tools and Pre-Service Teachers' Perspectives." *European Journal of Education* 60, no. 1: e12913.

Falala-Séchet, C., L. Antoine, I. Thiriez, and C. Bungener. 2019, July. "Owlie: A Chatbot That Provides Emotional Support for Coping With Psychological Difficulties." In Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents. 236–237. Paris, France.

Herbold, S., A. Hautli-Janisz, U. Heuer, Z. Kikteva, and A. Trautsch. 2023. "A Large-Scale Comparison of Human-Written Versus ChatGPT-Generated Essays." *Scientific Reports* 13, no. 1: 18617.

Hitsuwari, J., Y. Ueda, W. Yun, and M. Nomura. 2023. "Does Human–AI Collaboration Lead to More Creative Art? Aesthetic Evaluation of Human-Made and AI-Generated Haiku Poetry." *Computers in Human Behavior* 139: 107502.

Johansson, I. R. 2023. "A Tale of Two Texts, a Robot, and Authorship: A Comparison Between a Human-Written and a ChatGPT-Generated Text." https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1767661&dswid=-5415.

Landa-Blanco, M., M. A. Flores, and M. Mercado. 2023. "Human vs. AI Authorship: Does it Matter in Evaluating Creative Writing? A Pilot Study Using ChatGPT." https://doi.org/10.31234/osf.io/wjsm3.

Liu, W., and Y. Wang. 2024. "The Effects of Using AI Tools on Critical Thinking in English Literature Classes Among EFL Learners: An Intervention Study." *European Journal of Education* 59, no. 4: e12804. https://doi.org/10.1111/ejed.12804.

Merriam, S. B. 1998. *Qualitative Research and Case Study Applications in Education*. Revised and Expanded from Case Study Research in Education. San Francisco, CA: Jossey-Bass Publishers.

Mindner, L., T. Schlippe, and K. Schaaff. 2023. "Classification of Human-and AI-Generated Texts: Investigating Features for ChatGPT." In International Conference on Artificial Intelligence in Education Technology, 152–170. Singapore: Springer Nature Singapore.

Motlagh, N. Y., M. Khajavi, A. Sharifi, and M. Ahmadi. 2023. "The Impact of Artificial Intelligence on the Evolution of Digital Education: A Comparative Study of OpenAI Text Generation Tools Including ChatGPT, Bing Chat, Bard, and Ernie." arXiv preprint arXiv:2309.02029.

Myers, D., R. Mohawesh, V. I. Chellaboina, et al. 2024. "Foundation and Large Language Models: Fundamentals, Challenges, Opportunities, and Social Impacts." *Cluster Computing* 27, no. 1: 1–26.

Natekin, A., and A. Knoll. 2013. "Gradient Boosting Machines, a Tutorial." *Frontiers in Neurorobotics* 7: 21.

Ngo, G., R. Beard, and R. Chandra. 2022. "Evolutionary Bagging for Ensemble Learning." *Neurocomputing* 510: 1–14.

Ouyang, L., J. Wu, X. Jiang, et al. 2022. "Training Language Models to Follow Instructions With Human Feedback." *Advances in Neural Information Processing Systems* 35: 27730–27744.

Perera, P., and M. Lankathilake. 2023. "Preparing to Revolutionize Education With the Multi-Model GenAI Tool Google Gemini? A Journey Towards Effective Policy Making." *Journal of Advances in Education and Philosophy* 7: 246–253.

Perkins, M., J. Roe, D. Postma, J. McGaughran, and D. Hickerson. 2024. "Detection of GPT-4 Generated Text in Higher Education: Combining Academic Judgement and Software to Identify Generative AI Tool Misuse." *Journal of Academic Ethics* 22, no. 1: 89–113.

Salazar, G. Z., D. Zúñiga, C. L. Vindel, et al. 2023. "Efficacy of AI Chats to Determine an Emergency: A Comparison Between OpenAI's ChatGPT, Google Bard, and Microsoft Bing AI Chat." *Cureus* 15, no. 9: e45473. https://doi.org/10.7759/cureus.45473.

Sandag, G. A. 2020. "A Prediction Model of Company Health Using Bagging Classifier." *JITK* 6, no. 1: 41–46.

Schoppa, L., M. Disse, and S. Bachmair. 2020. "Evaluating the Performance of Random Forest for Large-Scale Flood Discharge Simulation." *Journal of Hydrology* 590: 125531.

Schweter, S. 2020. "BERTurk - BERT models for Turkish." https://doi.org/10.5281/zenodo.3770923.

Steiss, J., T. Tate, S. Graham, et al. 2024. "Comparing the Quality of Human and ChatGPT Feedback of Students' Writing." *Learning and Instruction* 91: 101894.

Van Dis, E. A., J. Bollen, W. Zuidema, R. VanRooij, and C. L. Bockting. 2023. "ChatGPT: Five Priorities for Research." *Nature* 614, no. 7947: 224–226.

Van Veen, D., C. Van Uden, L. Blankemeier, et al. 2023. "Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts." *Research Square*: rs.3.rs-3483777. arXivpreprint arXiv:2309.07430. https://www.nature.com/articles/s41591-024-02855-5.

Vaswani, A., N. Shazeer, N. Parmar, et al. 2017. "Attention Is All You Need." *Advances in Neural Information Processing Systems* 30: 5998–6008.

Wang, Y., Y. Xin, and L. Chen. 2024. "Navigating the Emotional Landscape: Insights Into Resilience, Engagement, and Burnout Among Chinese High School English as a Foreign Language Learners."

Learning and Motivation 86: 101978. https://doi.org/10.1016/j.lmot.2024.101978.

Wang, Y. S., and Y. X. Wang. 2024. "The 5Cs Positive Teacher Interpersonal Behaviors-Implications for Learner Empowerment and Learning in an L2 Context." *Porta Linguarum* 42: 363–366. https://doi.org/10.30827/portalin.vi42.29326.

Wu, H., Y. Wang, and Y. Wang. 2024. ""To Use or Not to Use?" A Mixed-Methods Study on the Determinants of EFL College Learners' Behavioral Intention to Use AI in the Distributed Learning Context." *International Review of Research in Open and Distributed Learning* 25, no. 3: 158–178. https://doi.org/10.19173/irrodl.v25i3.7708.

Yeşilçınar, S. 2023. "Personalized Learning Through Gamification: A ChatGPT Approach to English Language Learning." In *Transforming the Language Teaching Experience in the Age of AI*, 44–64. Hershey, PA: IGI Global. https://doi.org/10.4018/978-1-6684-9893-4.ch003.

Yıldırım, A., and H. Şimşek. 2013. *Sosyal bilimlerde nitel araştırma yöntemleri*. Ankara: Seçkin yayıncılık.

Yildiz Durak, H. 2023. "Conversational Agent-Based Guidance: Examining the Effect of Chatbot Usage Frequency and Satisfaction on Visual Design Self-Efficacy, Engagement, Satisfaction, and Learner Autonomy." *Education and Information Technologies* 28, no. 1: 471–488.

Yildiz Durak, H., and Ş. Aydoğdu. 2023. "Examining the Pandemic Process from the Educational Perspective: What Are Scientific Studies Telling Us?" *Journal of Learning and Teaching in Digital Age* 8, no. 2: 221–236.

Yildiz Durak, H., and A. Onan. 2023. "An Examination of Studies on the Use of Chatbot Technology in the Field of Education." *International Conference on Applied Engineering and Natural Sciences* 1, no. 1: 121–124.

Yildiz Durak, H., and A. Onan. 2024. "Predicting the Use of Chatbot Systems in Education: A Comparative Approach Using PLS-SEM and Machine Learning Algorithms." *Current Psychology* 43: 1–19.

Yin, R. K. 2003. "Designing Case Studies." *Qualitative Research Methods* 5, no. 14: 359–386.

Yin, R. K. 2009. *Case Study Research. Design and Method*. 4th ed. London: Sage Publications.

Ying, C., M. Qi-Guang, L. Jia-Chen, and G. Lin. 2013. "Advance and Prospects of AdaBoost Algorithm." *Acta Automatica Sinica* 39, no. 6: 745–758.

Yun-tao, Z., G. Ling, and W. Yong-cheng. 2005. "An Improved TF-IDF Approach for Text Classification." *Journal of Zhejiang University-Science A* 6, no. 1: 49–55.

Zhang, Y., and C. Zhang. 2024. "Extracting Problem and Method Sentence From Scientific Papers: A Context-Enhanced Transformer Using Formulaic Expression Desensitization." *Scientometrics* 129, no. 6: 3433–3468. https://doi.org/10.1007/s11192-024-05048-6.

Zhao, W., Y. Zhao, X. Lu, S. Wang, Y. Tong, and B. Qin. 2023. "Is ChatGPT Equipped With Emotional Dialogue Capabilities?" arXiv preprint arXiv:2304.09582.

Zhao, X., and Y. Wang. 2024. "EFL Teachers' Perceptions of Emotional Exhaustion and Associated Regulation Strategies: A Phenomenological Analysis." *Innovation in Language Learning and Teaching* 23: 1–15. https://doi.org/10.1080/17501229.2024.2391377.