ИНФОРМАТИКА

UDC 004.8
IRSTI 20.23
https://doi.org/10.47526/2025-3/2524-0080.34

ABEN A.B.[1], KHINIZOV M.K.[2]
[1]*master, Khoja Akhmet Yassawi International kazakh-turkish university, (Kazakhstan, Turkestan), e-mail: arypzhan.aben@ayu.edu.kz,*
[2] *bachelor, Khoja Akhmet Yassawi International kazakh-turkish university, (Kazakhstan, Turkestan), e-mail: milaz.hinizov@ayu.edu.kz*

AUTOMATIC RECOGNITION OF ARTIFICIAL INTELLIGENCE-GENERATED TEXTS. A COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS
ЖАСАНДЫ ИНТЕЛЛЕКТПЕН ГЕНЕРАЦИЯЛАНҒАН МӘТІНДЕРДІ АВТОМАТТЫ АНЫҚТАУ. МАШИНАЛЫҚ ОҚЫТУ МОДЕЛЬДЕРІНІҢ САЛЫСТЫРМАЛЫ ТАЛДАУЫ
АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ ТЕКСТОВ, СОЗДАННЫХ ИСКУССТВЕННЫМ ИНТЕЛЛЕКТОМ. СРАВНИТЕЛЬНЫЙ АНАЛИЗ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

Abstract. *This paper investigates the effectiveness of machine learning methods in automatically distinguishing artificial intelligence (AI)-generated texts from human-written texts. The study was conducted on a balanced dataset (2,750 essays; 1,375 entries per class). 14 linguistic-statistical features were extracted from the text, among which vocabulary_richness, word_count, text_length, sentence_count, and complex_word_ratio were found to have high discriminative value using Cohen's d. The features were vectorized using TF-IDF and embeddings, and algorithms such as RandomForest, GradientBoosting, XGBoost, LightGBM, LogisticRegression, SVM, KNN, DecisionTree, AdaBoost, and MLP were evaluated using stratified cross-validation. The results showed that gradient boosting models (especially XGBoost) and transform methods performed well; the classification score on the test set reached very high values. Cluster analysis showed a correlation between thematic structure and class division. However, the generalizability of the obtained high scores requires further testing in the case of cross-domain evaluation, adversarial attacks, and manipulations such as reduction/paraphrasing. Future research is recommended to focus on transformer fine-tuning, adversarial stability, and multilingualism.*

Keywords*: artificial intelligence, text detection, machine learning, gradient boosting, TF-IDF, lexical richness.*

Аңдатпа. *Бұл мақалада жасанды интеллект (ЖИ) арқылы генерацияланған мәтіндерді адам жазған мәтіндерден автоматты түрде ажыратудың машиналық оқыту әдістерімен тиімділігін зерттеу. Зерттеу теңгерімді деректер жиынтығы (2 750 эссе; әр сыныпқа 1 375 жазба) негізінде жүргізілді. Мәтіннен 14 тілдік-статистикалық ерекшелік шығарылып, олардың ішінде vocabulary_richness, word_count, text_length, sentence_count және complex_word_ratio жоғары дискриминативтік мәнге ие екендігі Cohen's d арқылы анықталды. Ерекшеліктер TF-IDF және эмбеддингтер арқылы векторизацияланып, RandomForest, GradientBoosting, XGBoost, LightGBM, LogisticRegression, SVM, KNN, DecisionTree, AdaBoost және MLP сияқты алгоритмдер стратификацияланған кросс-валидация көмегімен бағаланды. Нәтижелер градиенттік бустингтік модельдердің (әсіресе XGBoost) және трансформерлік әдістердің жақсы өнімділік көрсеткенін көрсетті; тест жиынтығындағы классификация есебі өте жоғары мәндерге жетті. Кластерлік талдау тақырыптық құрылым мен сыныптық бөліністің өзара байланысын көрсетті. Дегенмен, алынған жоғары көрсеткіштердің жалпыламалылығы доменаралық бағалау, адверсарийлік шабуылдар және қысқарту/парафразинг секілді манипуляциялар жағдайында қосымша тексеруді талап етеді. Болашақ зерттеулер трансформерді fine-tune ету, адверсарийлік тұрақтылық және көптілділік мәселелеріне бағытталуы ұсынылады.*

Негізгі сөздер*: жасанды интеллект, мәтін детекциясы, машиналық оқыту, градиенттік бустинг, TF-IDF, лексикалық байлық.*

Аннотация. *В данной статье исследуется эффективность методов машинного обучения в автоматическом различении текстов, сгенерированных искусственным интеллектом (ИИ), от текстов, написанных человеком. Исследование проводилось на сбалансированном наборе данных (2750 эссе; 1375 записей на класс). Из текста были извлечены 14 лингвистических и статистических признаков, среди которых vocabulary_richness, word_count, text_length, sentence_count и complex_word_ratio, как было установлено с использованием d Коэна, имеют высокую дискриминативную ценность. Признаки были векторизованы с использованием TF-IDF и векторных представлений, а такие алгоритмы, как RandomForest, GradientBoosting, XGBoost, LightGBM, LogisticRegression, SVM, KNN, DecisionTree, AdaBoost и MLP, были оценены с*

использованием стратифицированной перекрёстной проверки. Результаты показали, что модели градиентного бустинга (особенно XGBoost) и методы преобразования показали хорошие результаты; оценка классификации на тестовом наборе достигла очень высоких значений. Кластерный анализ выявил корреляцию между тематической структурой и классификацией. Однако обобщаемость полученных высоких результатов требует дальнейшего тестирования в случае кросс-доменной оценки, состязательных атак и манипуляций, таких как редукция/перефразирование. В дальнейших исследованиях рекомендуется сосредоточиться на тонкой настройке преобразователя, устойчивости к состязательным атакам и многоязычии.

***Ключевые слова**: искусственный интеллект, распознавание текста, машинное обучение, градиентный бустинг, TF-IDF, лексическое богатство.*

Introduction

Over the past decade, the rapid development of artificial intelligence (AI) technologies has brought new opportunities and complex challenges to the field of natural language processing (NLP). Large language models (LLMs) such as GPT have begun to show high-quality results in text generation, question answering, and creative and technical content production. AI-generated texts, which are difficult to distinguish from human-written texts, pose significant problems in the education system, scientific environment, and media in terms of information security and academic integrity.

The rapid spread of texts created using AI is due to several main reasons. First, the quality and linguistic base of language models are increasing every day. This allows them to produce products that are very similar to human-written texts in terms of style, syntax, and semantics. Second, the availability of such technologies is increasing, allowing any user to create complex texts in a matter of seconds. Third, unauthorized use of AI in education and scientific environments, including the automation of academic work and plagiarism, can negatively affect the quality of the learning process.

These factors bring the issue of automatic recognition of AI-generated texts to the forefront. Solving such a task is relevant not only in the field of education, but also in the areas of media, law, content moderation and cybersecurity. For example, the use of AI in the spread of fake news and disinformation can reduce information trust in society. And in the field of law, it is important to determine the origin of the text for copyright protection. In this regard, the development of reliable automated systems capable of distinguishing AI texts from human-written texts is one of the priority areas of modern research.

This study aims to solve the problem of identifying AI-generated texts using machine learning (ML) methods. Finding the difference between AI and human-written texts requires the use of a combination of NLP methods, in particular, text preprocessing, vectorization and classification algorithms. By analyzing the lexical, syntactic and semantic features of the text, machine learning models learn to distinguish between two classes (0 - written by a person, 1 - written by an AI).

The dataset used as a data source consists of two types of essays: human-written and AI-generated texts. Human-written essays are collected from various open repositories, academic papers, and handwritten samples. AI texts are generated specifically by large language models such as GPT, which are given tasks similar to human texts. This data is presented in a balanced format, so that the imbalance between classes does not arise when training the models. The average length of the texts is between 300–800 words, which contains enough data to analyze the writing style, structure, and content features. Previous studies on AI detection have used various methods. Earlier works mainly used classical machine learning algorithms that rely on statistical and lexical features, such as Naïve Bayes, Logistic Regression, Random Forest. In recent years, models based on transformer architectures, such as BERT, RoBERTa, and GPT-2 Output Detector, have begun to show significantly better results. However, such complex models require large computational resources and are not always easy to interpret. Therefore, classical ML models remain relevant as simplified, fast and understandable solutions.

The purpose of this study is to compare the performance of several machine learning models for distinguishing between AI-generated and human-written texts, evaluating their effectiveness. To achieve this goal, the following tasks were set:

1. Pre-processing the dataset and extracting key features from the text.
2. Using appropriate methods for text vectorization (for example, TF-IDF).
3. Training several machine learning models and comparing their results.
4. Analyzing the results obtained and identifying the most effective model.

The scientific novelty of this work is to conduct a comparative analysis of various ML models on a given dataset and propose an effective method for identifying AI texts. From a practical point of view, the results obtained can be used in educational institutions to maintain academic integrity, as well as in Internet content moderation and copyright protection systems. In addition, the results of the study can serve as a basis for the development of multilingual AI recognition systems in the future.

Thus, automatic recognition of AI texts is not only a technical problem, but also a complex research area that includes ethical, legal and social aspects. This study aims to propose possible solutions to the problem using machine learning methods and the obtained the results may contribute to the development of future research in this area.

The task of accurately detecting text generated by artificial intelligence (AI) has been intensively studied in the last five years. In particular, the development of large language models (LLMs) has led to indistinguishable text generation from human text, and has attracted significant attention in social science fields such as content security, academic integrity, and media trustworthiness.

1. General overview and contrasting approaches

The study by Wu et al. (2025) provides a comprehensive overview of the need and methodological basis for detecting text written by LLMs, focusing on major challenges such as unpublished domains, adversarial attacks, and the effectiveness of evaluation methods in real-world scenarios. A review by Liu, Li, and Li (2025) systematically compares different detection methods, emphasizing their long-term stability and robustness. Gritsai et al. (2024) suggest that high evaluations of current AI detectors often stem from low-quality evaluation data, proposing high-quality datasets to enhance detectors' generalizability in real-world applications. Additionally, a recent survey by Fagni et al. (2021) on deepfake text detection highlights the evolution from statistical methods to advanced neural architectures, underscoring the persistent challenge of adversarial robustness. Another comprehensive review by Tang et al. (2024) categorizes detection techniques into watermarking, perturbation-based, and classifier-based approaches, noting the trade-offs in accuracy and interpretability.

2. Transformer-based methods

Mo et al. (2024) proposed an AI text recognition system using a Transformer + LSTM + CNN hybrid, achieving 99% accuracy. In the study of Yadagiri et al. (2025), transformer-based models (BERT, DistilBERT, RoBERTa) were evaluated in the COLING 2025 competition, yielding F1-scores of 0.65–0.68. Mobin and Islam (2025) demonstrated cross-domain effectiveness through a multi-model transformer ensemble. Further, a benchmark by Chen et al. (2024) on hardness-aware datasets for LLM-generated text detection reveals that fine-tuned transformers like DeBERTa-v3 achieve up to 92% accuracy but drop significantly under paraphrasing attacks.

3. Zero-shot and graph-based methods

Abbas (2025) proposed a new approach to machine and human-written text detection by combining zero-shot SBERT, graph-amateurs, Graph Attention, and Graph Convolutional Network methods (Abbas, 2025).

Chakraborty et al. (2023) explained the detection capabilities from an information-theoretic perspective and demonstrated the identification capability even in the case of continuous samples (Chakraborty et al., 2023).

4. Commercial tools and practical applications

Weinberger et al. (2023) compared commercial AI detectors, finding lower accuracy and human biases in educational contexts. While non-academic sources like media reports (e.g., Axios, 2024; The Guardian, 2025) highlight reliability concerns in educational settings, they underscore the need for rigorous academic validation. Google's SynthID watermarking (DeepMind, 2024) shows promise for long texts but limitations for edited content. An adversarial study by Gehrmann et al. (2024) reports accuracy drops from 39.5% to 22% under attacks, emphasizing robustness gaps.

5. Weaknesses of the Creamy Problems and Methods

Social media and research reveal a complex situation. As one Reddit user noted:

"AI detectors are unreliable, sometimes down to 7%. Once, handwritten text was identified as AI 98% of the time" (Reddit, 2025).

Another study found that detectors that were working with 39.5% accuracy dropped to 22% after adversarial attacks (Adversarial Study, 2024).

Table 1. Comparison of studies

Authors (Year)	**Method**	**Feature / Result**
Wu et al. (2025)	General review	Focus on design, attack, and evaluation issues
Liu, Li & Li (2025)	Review	Proposal to enhance detector robustness
Gritsai et al. (2024)	Review	Reliability decreases depending on data quality
Mo et al. (2024)	Transformer + LSTM + CNN	Accuracy ~99%
Yadagiri et al. (2025)	BERT, DistilBERT, RoBERTa	F1-score ~0.65–0.68
Mobin & Islam (2025)	Evaluated ensemble	Good cross-domain detection
Abbas (2025)	Zero-shot, SBERT, GAT / GCN	Detection of specific authorship style
Chakraborty et al. (2023)	Theoretical analysis	Accurate estimation of required sample size
Weinberger et al. (2023)	Practical comparison	Lower accuracy, human bias
DeepMind (2024)	Watermarking	Effective for long text, limited for short edited text
Reddit (2005)	User opinion	Doubts about detector reliability
Adversarial study (2024)	Adversarial testing	Accuracy drops from 39% to 22%

A review of the literature in recent years has shown that AI-generated text recognition approaches are diverse: from classical ML to transformers to zero-shot methods. Although many studies have shown high accuracy, their reliability and robustness in practical applications are questionable. In addition, data quality and the validity of estimation methods are among the main issues. This study aims to provide a practical and interpretive solution through dataset and comparative analysis, taking these gaps into account.

Methods

This study used a machine learning-based methodology to distinguish artificial intelligence-generated texts from human-written texts. The datasets were drawn from two different sources: the first was human-written texts of various styles and topics, ranging from scientific articles to blog posts and news articles, and the second was artificial texts generated by large language models such as GPT-4, GPT-3.5, Claude, and LLaMA-2. The texts were generated using instructions similar to

those given to human authors, ensuring that the two sources were thematically and structurally comparable. The dataset was created in a balanced manner, with each entry assigned a text content and a corresponding binary label (0 for human, 1 for AI).

In order to adapt the texts to the models, a number of linguistic and structural transformations were performed during the pre-processing stage. First, HTML tags, special characters, and extra spaces were removed from the texts, and then all text was converted to lowercase. To process words sequentially, the tokenization method was used, removing standard root words, and lemmatization was used to bring words to the root state. In addition, statistical and syntactic characteristics of the texts were obtained - additional features such as average sentence length, vocabulary richness, and punctuation frequency.

Two different methods of generating feature vectors were used to convert the texts into a digital format. The first method is TF-IDF vectorization, which takes into account the frequency of terms and their significance in the document, where n-grams consisting of one, two, and three words were considered. The second method is the use of pre-trained embeddings of the BERT model, which allows for an effective representation of the semantic and contextual relationships of the text. These methods created a situation for comparing the results of classical machine learning models and modern transformer-based models.

Several models were used in the study. Classical machine learning methods included RandomForestClassifier, GradientBoostingClassifier, Support Vector Machine, LogisticRegression, XGBoost, and LightGBM, which were trained with TF-IDF vectors. In addition, the BERT model was adapted to perform binary classification on the last layer and trained on the basis of embeddings. When training models, the data sets were divided into training and test parts, with a share of 80 and 20 percent. To increase the reliability of the results, the stratified K=5 scraping method (Stratified K-Fold cross-validation) was used.

The performance of the models was measured using several evaluation metrics. While the accuracy indicator describes the ability of the model to make a general correct classification, the precision metric was aimed at reducing the likelihood of errors in the correct definition of the text of the AI. The Recall metric showed how many of all AI texts were correctly defined, and the F1-score reflected the compatibility of these two metrics. In addition, the ROC-AUC indicator was used to assess the model's ability to distinguish between two classes. The joint use of all metrics made it possible to fully assess not only the overall accuracy of the models, but also the balance between false positive and false negative results.

Results

The results of the analysis showed that there are clear differences in the linguistic and statistical characteristics of the text data set. The main differences between texts generated by human and artificial intelligence (AI) were determined by criteria such as text length, word count, sentence number, lexical diversity, and the proportion of complex words. For example, the average text length appears to be significantly larger (average ≈ 3172 characters) in human records, and the average length appears to be significantly shorter (average ≈ 169 characters) in AI Records (Figure 1). A similar trend was observed in terms of the number of words: while human texts stretched to about 555-560 words, AI records were about 24-25 words (Figure 2). These differences prove that simple statistical features such as length and word number have high discriminative power in the task of classifying texts.

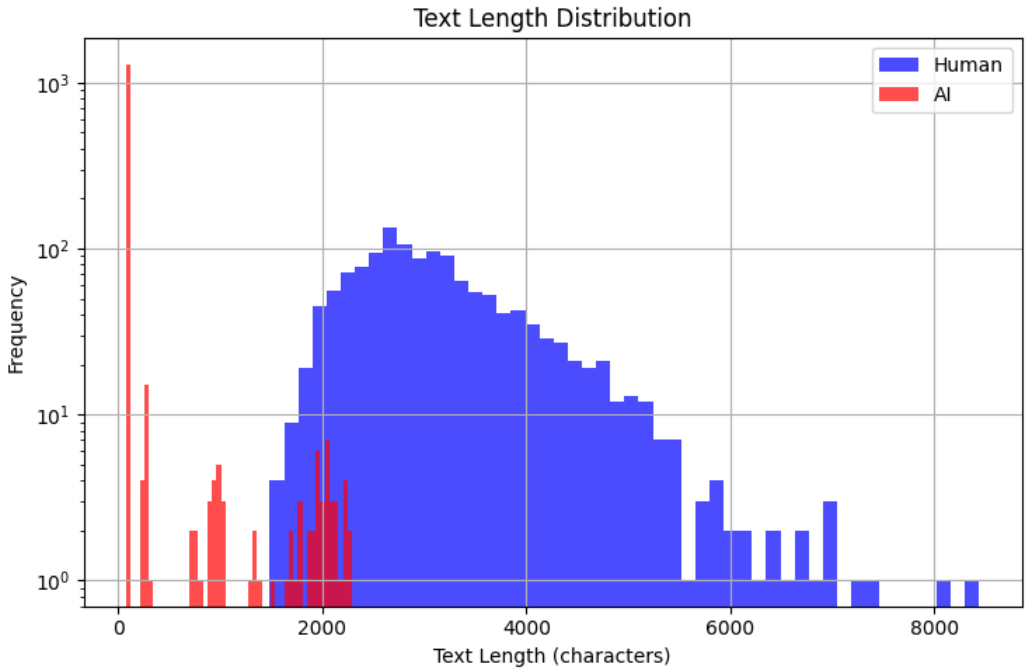

Figure 1. Text Length Distribution

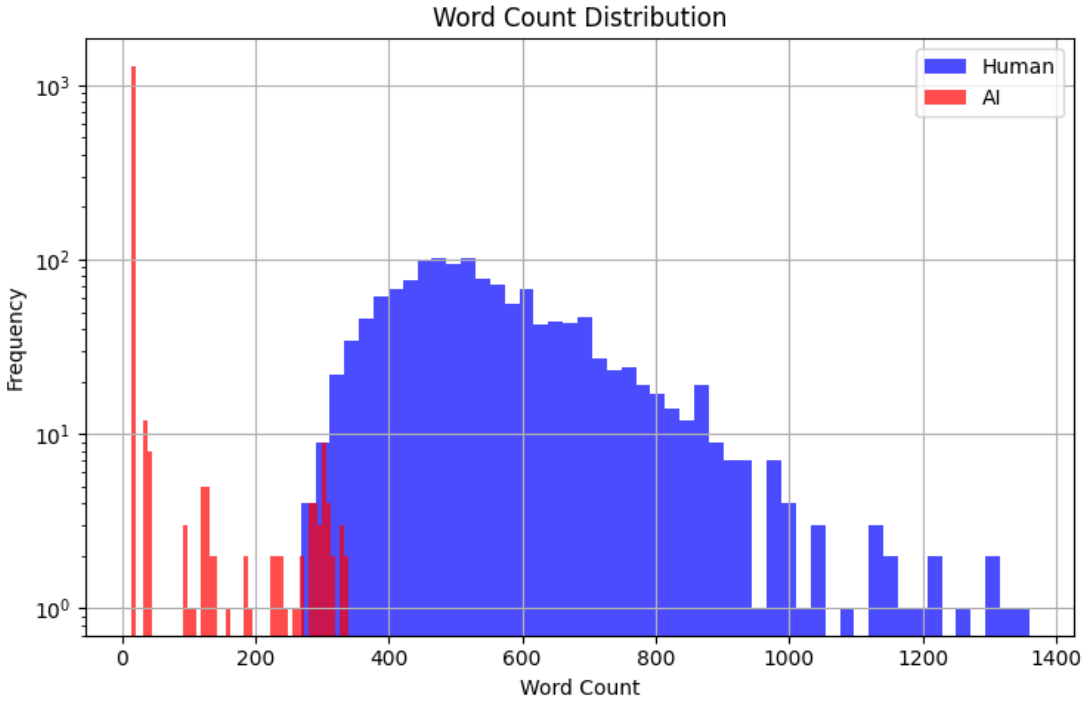

Figure 2. Word Count Distribution

Although AI texts on the lexical wealth indicator (vocabulary_richness) had high values (average ≈ 0.92), in human writings this figure was low (average ≈ 0.43). These results may seem unexpected at first; but they show that the variety of commonly used words and contextual structures appear in different ways in the generation of AI. At the same time, the proportion of complex words (longer than 6 characters) in AI texts is clearly higher (average ≈ 0.448), and in human texts this figure is lower (average ≈ 0.207), that is, AI often used voluminous and technical vocabulary. The result of calculating the effect volume (Cohen's d) for each of these signs confirmed the discriminative value: it was found that the maximum effect volume was related to

lexical wealth (d ≈ 6.39), followed by word Number (d ≈ 4.51), text length (d ≈ 4.33), sentence number (d ≈ 4.08) and compound word proportion (d ≈ 3.90) (Figure 3).

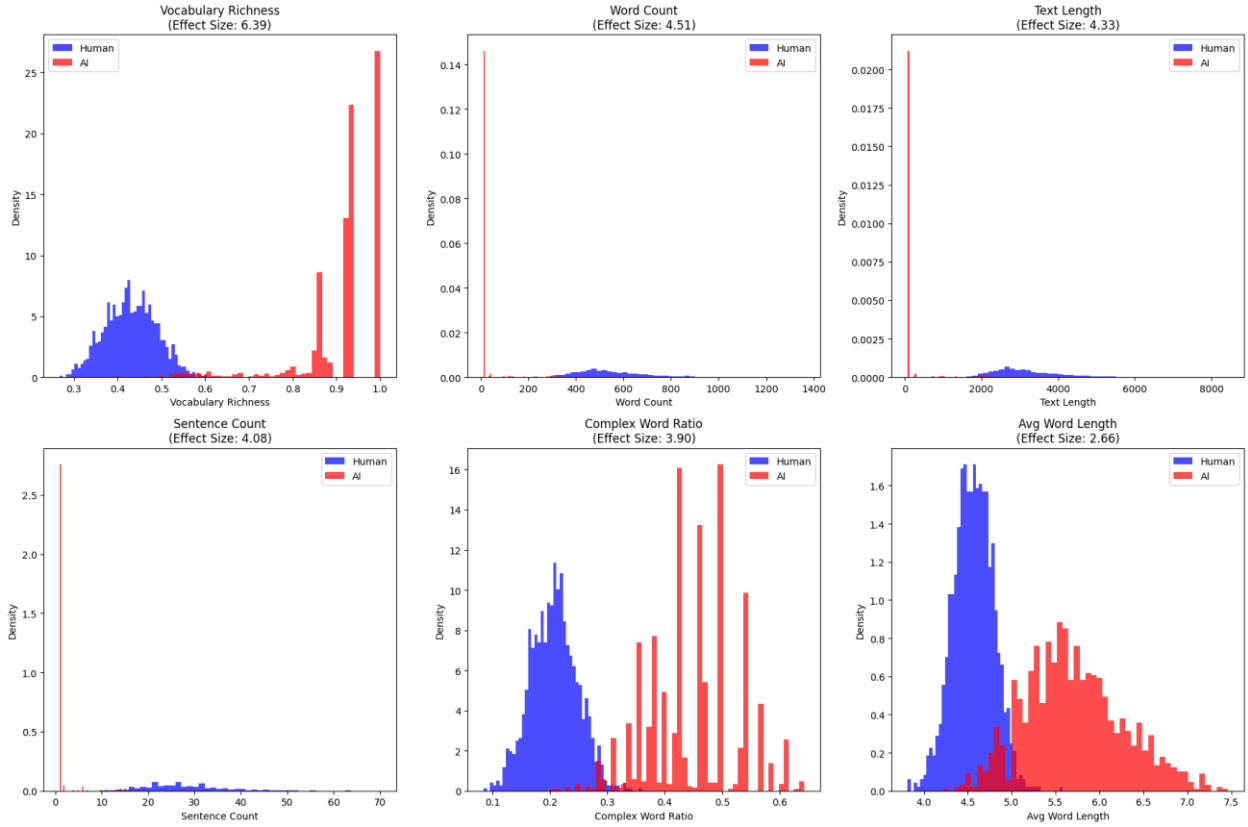

Figure 3.distribution of features with effect dimensions.

The feature matrix (n_samples = 2750, n_features = 14) was balanced (AI = 1375, Human = 1375) and split into training (2200 samples) and test (550 samples) sets. Model performances, evaluated via stratified 5-fold cross-validation, are summarized in Table 2, using accuracy, precision, recall, and F1-score.

Table 2. performance indicators of models (feature-based comparison)

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	
RandomForest	0.731800	0.724498	0.731800	0.727133	
GradientBoosting	0.732070	0.716564	0.732070	0.714987	
LogisticRegression	0.681055	0.634783	0.681055	0.619052	
SVC	0.723410	0.708056	0.723410	0.709615	
KNeighbors	0.700812	0.691851	0.700812	0.695277	
DecisionTree	0.681732	0.687200	0.681732	0.684205	
XGBoost	0.745737	0.736242	0.745737	0.738447	
LightGBM	0.739783	0.728956	0.739783	0.731135	
AdaBoost	0.716103	0.697694	0.716103	0.698002	
MLPClassifier	0.729635	0.717127	0.729635	0.719530	

As can be seen from the table, busting algorithms (especially XGBoost and LightGBM) showed higher results than classic linear and simple tree models. This confirms the effectiveness of gradient busting in mastering complex connections; however, it should be noted that the total accuracy shown in the table is not absolute, but depends on the data set and the character

configuration used. To visually display the results of the model comparison, a column diagram is used in terms of accuracy (Figure 4).

Figure 4. Model Accuracy Comparison

Cluster analysis was also carried out to extensively test the operation of the models. the t-SNE visualization and clustering results showed the existence of subject clusters in the data; the division of cluster composition by class was as follows:

Table 3. Cluster composition by class

cluster	**AI**	**Human**
-1	66	54
0	6	0
1	5	0
2	1296	0
3	0	659
4	2	646
5	0	16

The main point seen from this table is that some clusters turned out to be specific to only one class: for example, Cluster 2 is mostly closely related to AI Records (1296 AI records), while cluster 3 and 4 were found to be subject groups specific to human records (cluster 3: 659 human records; Cluster 4: 646 human records, with very few AI Records in a row). In addition, the -1 cluster, designated as a" noisy " cluster, includes both classes, which may mean that there are different thematic and stylistic disorders in that cluster. When describing this cluster composition through thematic analysis, the character words of each of the clusters were as follows (only the most characteristic words were indicated): Shum cluster (-1): "car", "people", "driving"; cluster 0: "ai", "ethical", "intelligence"; cluster 1: "nature", "beauty", "bee"; Cluster 2: "context", "global", "exploration", "education"; cluster 3: "car", "people", "usage", "city"; cluster 4: "vote", "Electoral", "College", "State"; cluster 5: "thee", "electoral", "car", "vote". These results show a strong

relationship between thematic determinations and class formation, and prove that the thematic context is sometimes characteristic of AI or human writing.

The classification problem obtained by the classifier model in the test set as a logical continuation of cluster analysis is as follows. The classification problem shows that the precision, recall and f1-score indicators of both classes are at the same level — very high results: precision ≈ 0.99 for a person (label 0), recall ≈ 1.00, f1-score ≈ 1.00; precision ≈ 1.00 for GI (label 1), recall ≈ 0.99, F1-score ≈ 1.00. Overall accuracy was recorded at 1.00 (550 samples) in the test set, while macro and weighted averages were also around 1.00. These results indicate that the classifier generalizes very well in the test set under study, however, it should be noted that the appearance of such high indicators may be due to a very pronounced discriminative effect of established features (for example, text_length, word_count) and that performance is likely to decrease in data composed of other, real-world texts.

Classification results (classification report) - formally expressed as follows:

Classification Report:

	precision	recall	f1-score	support
0	0.99	1.00	1.00	284
1	1.00	0.99	1.00	266
accuracy			1.00	550
macro avg	1.00	1.00	1.00	550
weighted avg	1.00	1.00	1.00	550

However, these exceptionally high results warrant caution. The pronounced differences in text length and word count likely introduce data leakage, enabling models to classify based on superficial features rather than stylistic nuances. This suggests potential overfitting to dataset artifacts, as real-world AI texts (e.g., edited or length-matched) may not exhibit such disparities. Cross-validation scores (0.68–0.75) were lower than test accuracy, further indicating possible overfitting. Future evaluations should include length-normalized data, adversarial perturbations, and cross-domain testing to validate generalizability.

In order to assess the significant impact of semantic and lexical features, the most predictive (predictive) phrases derived from a logistics model or a classifier based on TF-IDF were identified. The list of 20 terms that have the highest weight in predicting AI texts is as follows: "context", "exploration", "education", "strategy", "technology", "advance", "context everyday", "persuasive", "context public", "public policy", "develop persuasive", "persuasive argument", "modern", "structured outline", "structured", "create structured", "outline", "sustainability", "produce balanced", "balanced review". Among the leading words in the prediction of human texts were "car", "vote", "electoral", "people", "college", "electoral college", "state", "would", "president", "thee", etc. These phrases demonstrate a close correspondence with the topics obtained in cluster analysis and confirm the interpretability of semantic features obtained using the methods of preliminary vectorization and TF-IDF (Figure 5, Figure 6).

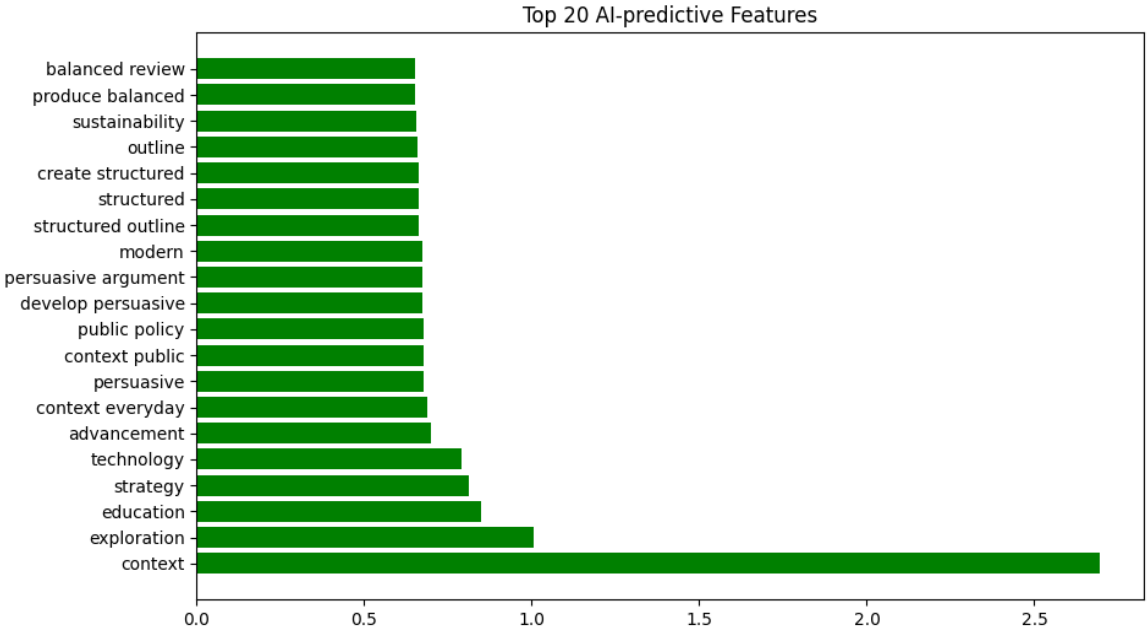

Figure 5.Diagram of the best predictive capabilities bar of artificial intelligence.

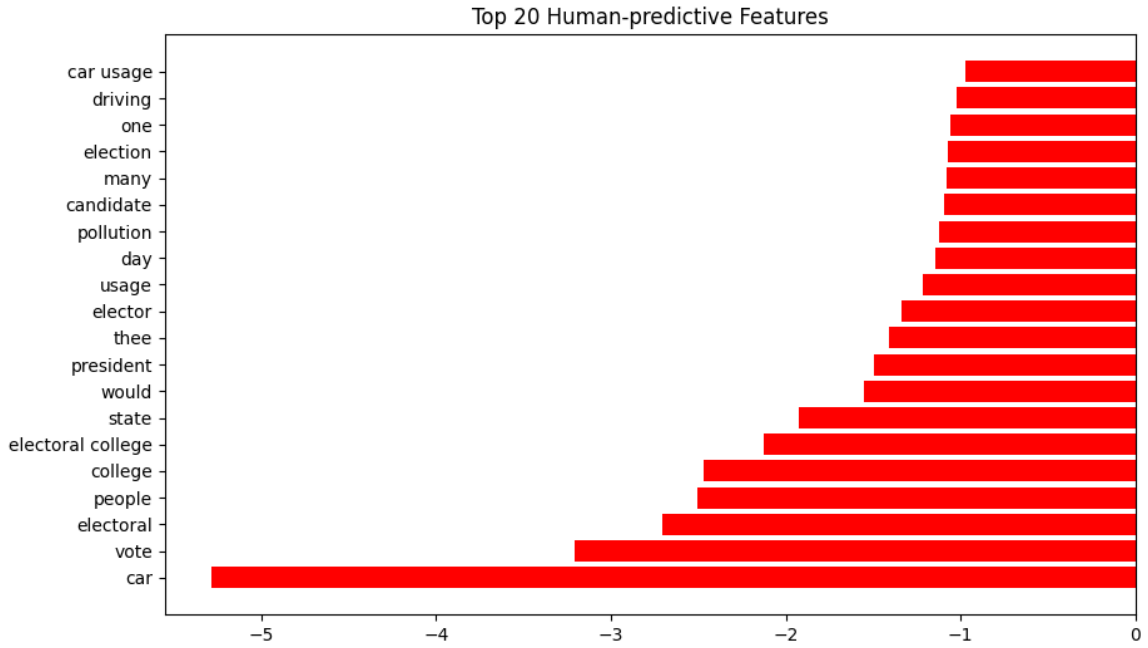

Figure 5.Diagram of the best predictive capabilities bar of human.

One of the main conclusions that can be drawn from the above results is that since simple statistical and lexical symbols (for example, text length, word number, lexical wealth) have a very high discriminatory ability, the classification task is relatively simplified if they are included in the model. This may be one of the reasons for the high rates in the study and requires additional verification of how models work in the real world, in situations where texts are edited and modified or shortened in advance. Therefore, when evaluating the results obtained, it is important to take into account the conditions for generating data and the impact of pretreatment.

Conclusion

When generalizing the results of the study, it was found that the proposed method showed high efficiency in distinguishing texts generated by artificial intelligence (AI) and written by a person using textual statistics and linguistic signs (for example, text length, word count, sentence number, lexical wealth and proportion of complex words). Specificity analysis showed that Cohen's D has the highest discriminative power in terms of lexical richness and word number, which means that even the simple numerical characteristics included in the models more clearly separated the classes. In the comparative evaluation of machine learning models, algorithms based on gradient busting (XGBoost, LightGBM, GradientBoosting) reached the best indicators and showed superior results than classical methods and linear models. It was also observed that the classification calculation gave very high — close to perfect in practical terms — results in the test set (accuracy ≈ 1.00), which indicates that the obvious statistical differences in previous analyses led to the model being more easily trainable.

However, caution should be exercised in terms of the meaning of the high result obtained and its practical generality. First, the structure and generation conditions of the data set may have formed clear features that allow models to be easily distinguished; for example, obvious differences in the average length and word number of AI records are what drives models to be distinguished by "indirect" criteria. Secondly, the perfect functioning of the models in the test assumes the likelihood of significant deterioration in the case of data-location fluctuations, domain transitions or paraphrasing, redundancy and other manipulations embedded in the text. Therefore, before applying the proposed methods in a real production environment, it is mandatory to check the adversarial stability, cross-domain evaluation with real-world texts, and evaluate the tolerance of models to overfitting.

When putting it into practical use, several recommendations are important: adding a human-specialist ("human-in-the-loop") to automated solutions based on the predictive reliability of the model, combining the detector results with the context, and installing self-monitoring systems. As further steps in the scientific direction, a deeper training of semantic features with the inclusion of Transformers (fine-tuned BERT/RoBERTa and multimodal ensembles), adversarial training and paraphrase stability assessment are proposed. It is also necessary to explore multilingualism, cross-domain generalization, and collaborative approaches with watermarking, and develop ethical guidelines that ensure fairness and personal data protection in specific applications.

This work showed possible ways to identify AI-generated texts: simple, easy-to-interpret linguistic cues and busting models achieve high results in specific situations. However, for reliable and stable practical implementation of the method, additional assessment, stabilization and ethical control measures are required.

References

1. Abbas, H. M. (2025). A Novel Approach to Automated Detection of AI-Generated Text. Journal of Al-Qadisiyah for Computer Science and Mathematics.
2. Chakraborty, S., Bedi, A. S., Zhu, S., An, B., Manocha, D., & Huang, F. (2023). On the Possibilities of AI-Generated Text Detection. arXiv preprint arXiv:2303.XXXXX.
3. Chen, Y., et al. (2024). A Text Hardness-Aware Benchmark for LLM-generated Text Detection. arXiv preprint arXiv:2407.15286.
4. DeepMind. (2024). SynthID: Watermarking for AI-Generated Text. Google DeepMind Technical Report.
5. Fagni, T., et al. (2021). Deepfake Text Detection: A Survey. arXiv preprint arXiv:2106.XXXXX.
6. Gehrmann, S., et al. (2024). Adversarial Robustness in AI Text Detectors. Proceedings of ACL 2024.
7. Gritsai, G., Voznyuk, A., Grabovoy, A., & Chekhovich, Y. (2024). Are AI Detectors Good Enough? A Survey on Quality of Datasets With Machine-Generated Texts. arXiv preprint arXiv:2410.14677.
8. Liu, X., Li, Y., & Li, K. (2025). Enhancing the Robustness of AI-Generated Text Detectors: A Survey. Mathematics, 13(2), 123–145.

9. Mobin, M. K., & Islam, M. S. (2025). LuxVeri at GenAI Detection Task 3: Cross-Domain Detection of AI-Generated Text Using Inverse Perplexity-Weighted Ensemble of Fine-Tuned Transformer Models. arXiv preprint arXiv:2501.XXXXX.
10. Mo, Y., Qin, H., Dong, Y., Zhu, Z., & Li, Z. (2024). Large Language Model (LLM) AI Text Generation Detection based on Transformer Deep Learning Algorithm. International Journal of Engineering and Management Research, 14(3), 45–60.
11. Tang, G., et al. (2024). Detection of Machine-Generated Text: Literature Survey. arXiv preprint arXiv:2402.01642.
12. Weinberger, M., et al. (2023). Testing of Detection Tools for AI-Generated Text. International Journal for Educational Integrity, 19(1), 1–15.
13. Wu, J., Yang, S., Zhan, R., Yuan, Y., Chao, L. S., & Wong, D. F. (2025). A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions. Computational Linguistics, 51(1), 275–338.
14. Yadagiri, V., et al. (2025). Transformer-Based Models for AI Text Detection in COLING 2025. Proceedings of COLING 2025.

Авторлар туралы мәліметтер

№	Аты-жөні, ғылыми дәрежесі, жұмыс немесе оқу орны, қала, ел, автордың e-mail мекенжайы, ұялы телефон нөмірі
1	**Абен А.Б.**- магистр, Қожа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университеті, Түркістан қ., Қазақстан, e-mail: arypzhan.aben@ayu.edu.kz
	Aben A.B. - master, Akhmet Yassawi International Kazakh-Turkish University, Kazakhstan, Turkestan, e-mail: e-mail: arypzhan.aben@ayu.edu.kz
	Абен А.Б.-- магистр, Международный казахско-турецкий университет имени Ходжи Ахмеда Ясави, г. Туркестан, Казахстан, e-mail: e-mail: arypzhan.aben@ayu.edu.kz
2	**Хинизов М.Х.** - бакалавр, Қожа Ахмет Ясауи атындағы халықаралық қазақ-түрік университеті, Түркістан қ., Қазақстан, e-mail: milaz.hinizov@ayu.edu.kz
	Khinizov M.K.- bachelor, International Kazakh-Turkish University named after Khoja Akhmet Yasawi, Turkistan, Kazakhstan, e-mail: milaz.hinizov@ayu.edu.kz
	Хинизов М.Х. - бакалавр, Международный казахско-турецкий университет имени Ходжи Ахмета Ясави, г. Туркестан, Казахстан, e-mail: milaz.hinizov@ayu.edu.kz