КОМПЬЮТЕРНЫЕ НАУКИ, ПРИБОСТРОЕНИЕ И АВТОМАТИЗАЦИЯ
КОМПЬЮТЕРЛІК ҒЫЛЫМДАР, АСПАП ЖАСАУ ЖӘНЕ АВТОМАТТАНДЫРУ
COMPUTER SCIENCE, INSTRUMENTATION AND AUTOMATION

UDC 004.04
IRSTI 28.23.25
https://doi.org/10.53364/24138614_2025_39_4_13

A.A. Sagynbayev[1*], D.Z. Kaibassova[1], A.A. Sagynbayeva[1]
[1]Astana IT University, Астана, Казахстан

[*]*E-mail: 232040@astanait.edu.kz*

COMPARATIVE ANALYSIS OF AI-GENERATED TEXT DETECTION MODELS IN STUDENT TEXTUAL ASSIGNMENTS

Abstract. *Modern transformer models have significantly expanded the capabilities of automated text generation, posing new challenges for maintaining academic integrity in higher education. Traditional plagiarism detection systems often fail to distinguish between student-written work and AI-generated materials, underscoring the need for robust automatic detectors. Accordingly, this article presents a comparative analysis of three approaches to detecting AI-generated text in student submissions. The study focuses on a GPT-2 transformer-based classifier, a CNN-LSTM hybrid architecture, and a classic LSTM model. Research objectives include standardizing an experimental protocol and evaluating each method under varying computational constraints and accuracy requirements. The experimental methodology comprises uniform preprocessing of a labeled corpus of student assignments, splitting data into training and validation sets, training models over multiple epochs with identical tokenization and optimization parameters, and assessing their performance using precision, recall, and F1-score metrics. Findings reveal that the transformer-based detector provides the deepest contextual representations, the CNN-LSTM hybrid achieves an optimal balance between processing speed and detection quality, and the LSTM model serves as an efficient, resource-saving baseline for CPU-only environments. The authors conclude that method selection should align with available infrastructure: transformers are suited for GPU-rich servers, hybrid architectures for mid-range platforms, and LSTM modules for CPU-based setups. As a practical recommendation, the authors propose integrating the hybrid detector into educational platforms alongside expert peer review and regularly updating the training corpus to adapt to emerging types of AI-generated content.*

Keywords: *Deep Learning, AI-Generated Text, Academic Integrity, GPT-2, CNN-LSTM, LSTM, Text Classification, Transformer Models, Hybrid Models.*

Introduction.

Recent advancements in natural language processing (NLP) have led to the development of advanced language models that can generate text nearly resembling human writing. Content creation in many different fields has been revolutionized by transformer-based approaches such GPT and BERT. However, these developments pose significant challenges to academic integrity in higher education, where separating AI-generated from human-written materials is becoming more and more important [1]. Mostly depending on apparent textual similarities, traditional plagiarism detection algorithms sometimes fail to identify the subtle outputs produced by modern artificial intelligence systems [2].

Deep learning-based approaches have been explored to find methods of handling these difficulties. Since Long Short-Term Memory (LSTM) networks are excellent in capturing long-range dependencies in text, they are ideal for modeling sequential patterns. LSTMs could thus

overlook minor local traits that distinguish actual content from machine-generated text. Researchers have suggested hybrid models mixing LSTM architectures with CNNs to address this. Combining the LSTM's sequential modeling capability with CNN's mastery in local feature extraction, these CNN-LSTM hybrids increase detection accuracy [3, 4].

Moreover, transformer-based models meant for text generation have been modified to recognize tasks. By fine-tuning models like GPT-2 for classification or by looking at perplexity measurements [5], researchers have found unusual patterns indicating of AI-generated text. Every method has special advantages and drawbacks for scalability, computing economy, and accuracy.

The reason for implementing advanced models is clear: as higher education confronts the rising challenge of academic dishonesty through AI-generated submissions, there is an urgent need for effective automated detection methods. These technologies preserve academic integrity and enhance digital content management on educational platforms [1, 2]. This study performs a comparative analysis of three detection methodologies: LSTM-based, CNN-LSTM hybrid, and GPT-2-based approaches, to assess their effectiveness in identifying AI-generated text and its potential applications in higher education.

Literature Review.

For text classification, deep learning approaches have been looked into in great detail; researchers have investigated several architectures to address problems in identifying traditional and artificial-generated content. Especially LSTM and GRU variants, preliminary comparative studies [6] provide a comprehensive evaluation of architectures including Convolutional Neural Networks (CNN), Deep Belief Networks (DBN), and Recurrent Neural Networks (RNN), across many classification tasks. Their analysis shows that whereas CNNs are particularly adept at extracting local, position-invariant characteristics, RNN-based models are competent in capturing long-range dependencies. Recent studies underscore the efficiency of LSTM-based models in environments with limited computational resources, showing approximately 97 % accuracy in detecting AI-generated texts even in languages with limited datasets. Given that performance is quite sensitive to configuration selections, the study emphasizes the fundamental relevance of hyperparameter tuning. Nonetheless, these models are less effective for shorter academic texts, where local and nuanced textual features play a significant role, thereby limiting their applicability in certain educational settings [7].

Building upon this basis, a hybrid CNN-LSTM model was proposed for the detection of fake news [8]. This method utilizes the CNN component to extract prominent textual elements, while the LSTM layer identifies the sequential relationships essential for differentiating between authentic and fabricated news. The hybrid model demonstrated exceptional performance on the ISOT Fake News Dataset, surpassing both conventional classifiers and independent deep learning models. Recent comparative analyses demonstrate hybrid CNN-LSTM models' superior accuracy (up to 99 %) due to their balanced integration of local textual features captured by CNNs and sequential context modeled by LSTMs. However, these hybrid architectures often require considerable computational power and careful optimization to avoid overfitting, posing potential constraints for institutional use [4]. This study emphasizes the benefit of integrating complementary architectures to overcome individual limitations.

Additionally, recent studies have focused on transformer-based models to improve text categorization and the identification of AI-generated material. One investigation examined the use of BERT through its fine-tuning on datasets containing both human-written and AI-generated text [9]. The research demonstrates that BERT, utilizing transfer learning with limited labeled data, attains elevated accuracy, precision, and recall across several domains. The feature significance analysis indicates that contextual embeddings obtained from BERT's attention processes are crucial for differentiating AI-generated material from human-written language. Transformer-based detection methods achieve robust F1-scores (around 90 %) thanks to their advanced contextual representations and self-attention mechanisms.

Further extending the discussion on advanced language models, GPT-2 was introduced as a transformative, unsupervised multitask learner [10]. Trained on the WebText dataset in a zero-shot setting, GPT-2 exhibits competitive performance across numerous NLP benchmarks, including question answering, machine translation, and summarization. The research indicates that GPT-2's performance increases log-linearly with model size, highlighting the potential of large-scale transformer models. Though originally intended for text production, the GPT-2 insights are especially important for spotting AI-generated text. Its ability to create coherent, human-like language not only makes it more difficult to distinguish between human- and artificial-generated content but also provides a basis for creating more advanced detection systems. Yet, these models also exhibit vulnerabilities, such as high computational demands, sensitivity to slight textual alterations, and dependency on extensive fine-tuning data, thus potentially limiting their practical use across varied academic contexts [5,11]. Moreover, these models can be used to improve automated grading, essay evaluation, and plagiarism detection systems, so affecting higher education as well. Table 1 summarizes a SWOT analysis of the three major classes of AI-generated-text detectors — GPT-2-based, CNN-LSTM hybrid и LSTM-based architectures — highlighting their relative strengths, weaknesses, opportunities и threats.

Table 1 – SWOT Analysis of AI-Generated Text Detection Models

Model	Strengths	Weaknesses	Opportunities	Threats
LSTM - Based	Lightweight, efficient sequential modeling	Limited local feature detection; struggles with short texts	Leverage in low-resource environments; easy deployment on CPU-only systems.	Advanced adversarial text generation may exploit sequential gaps.
CNN - LSTM Hybrid	Combines local feature extraction with temporal context	Increased model complexity; higher training cost	Hybrid architectures can be tuned for optimal trade-off; supports on-premise use	Overfitting on small datasets; maintenance overhead.
GPT - 2 Based	Superior contextual embeddings, highest accuracy	Resource-intensive; sensitive to slight text variations	Regulatory frameworks encourage adoption; fine-tuned domain models.	Rapid evolution of LLMs may outpace detector updates.

Apart from these deep learning approaches, traditional machine learning methods have also been applied for evaluation of academic texts. To assess the structure and formatting quality of student articles, a comparison study [12] including k-nearest neighbors, support vector regression, and random forest was conducted on models. Their results highlight how conventional machine learning techniques could evaluate text quality, so offering a different perspective on the more recent deep learning techniques meant for material generated by artificial intelligence.

Additionally, the rise of advanced artificial intelligence-generated writing gradually compromises academic integrity in higher education. A recent study looks at the complicated problem of plagiarism and proposes that developing technology—especially artificial intelligence-based detection systems—may greatly help to prevent academic misbehavior [13]. The study emphasizes the need of advanced, technologically developed detection methods to support institutional projects maintaining ethical standards by means of technology.

Recent developments in the US and Europe show widespread adoption of AI-generated text detection within academic integrity policies, bolstered by the EU’s Artificial Intelligence Act’s emphasis on transparency and ethical AI governance. In Kazakhstan and other Central Asian countries, universities are adapting international best practices to manage generative AI despite

the absence of a unified national framework. Across broader Asia, institutions favor balanced, cautious approaches that pair automated detection with human review to reduce errors and bias. Frameworks like the AI Ecological Education Policy Framework explicitly address pedagogical, operational, and governance dimensions of responsible AI use in teaching [14], while the Higher Education Act for AI (HEAT-AI) offers a risk-based regulatory model tailored for HEIs to ensure accountability in detecting AI-generated content [15]. Effective integration in settings such as Kazakhstan therefore requires combining advanced detection tools, clear institutional guidelines, pedagogical adjustments, and vigilant human oversight to uphold academic integrity without unfairly penalizing students.

These studies all together demonstrate the progression of text classification methodologies, transitioning from initial architectures highlighting either local feature extraction or sequential modeling to hybrid approaches that blend these functions, finishing in transformer-based methods that represent intricate contextual relationships. The consequences for higher education are important: the great efficiency of these models provides a good basis for the development of automated verification systems as academic institutions face the difficulty of spotting submissions created by artificial intelligence and maintaining academic integrity. Nonetheless, important domains for ongoing study are dataset variety, model generalization, and the use of new transformer topologies.

Materials and research methods.

Data Description

The "LLM - Detect AI Generated Text Dataset" [16] was chosen for this research and was obtained from Kaggle. Academic essays categorized as either human-written or AI-generated make up the dataset. To standardize the inputs, raw text samples were first preprocessed—that is, converted to lowercase and punctuation removed. Using an 80/20 ratio, the dataset was split into training and validation subsets thereby preserving balanced class distributions in both sets. To guarantee consistency in input dimensions, uniform preprocessing—including tokenization, normalization and padding to a set length of 512 tokens with PyTorch's pad_sequence—was applied across all models.

Model Architectures

The first method utilizes a standard LSTM-based model for identifying AI-generated text. As seen in Figure 1, the input goes through preprocessing, before its transformation into a sequence of token indices. These indices are then passed to a 128-dimensional embedding layer, which converts them into dense vector representations. A bidirectional LSTM with two layers and a hidden size of 256 subsequently handles the embeddings. The output from the final time step of the LSTM is processed through a dropout layer and subsequently directed into a fully connected classifier, yielding a single logit for binary classification. This model utilizes extensive sequential data inside the text, rendering it proficient at differentiating between human-generated and AI-generated content.

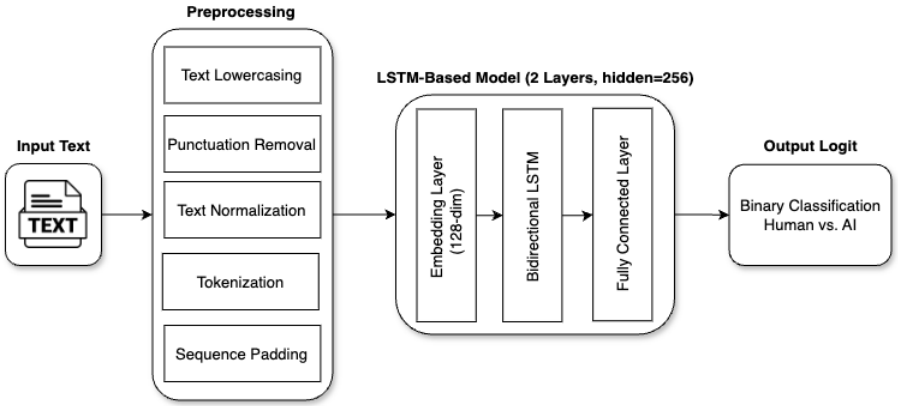

Figure 1 – LSTM-based detection model

The second method employs a hybrid model that integrates convolutional and LSTM layers to capture both local and long-range dependencies in text. As shown in Figure 2, the preprocessing steps, are similar to those of the LSTM-based method. Post-preprocessing, token indices are converted into 128-dimensional embeddings, subsequently processed through a 1D convolutional layer featuring 100 filters and a kernel size of 3 to extract local features. A ReLU activation and dropout layer follow, with the convolutional outputs reshaped to serve as input for a bidirectional LSTM with a hidden dimension of 128. The final forward and backward hidden states are concatenated and directed to a fully connected layer, producing a single logit for binary classification. By combining CNN-based feature extraction with LSTM-based temporal modeling, this architecture efficiently differentiates human-authored text from AI-generated content.

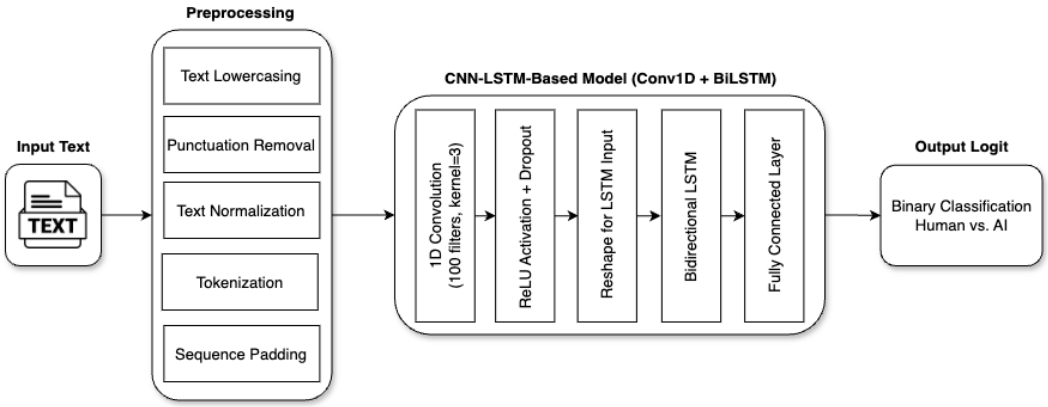

Figure 2 – CNN-LSTM-based hybrid detection model

The final method utilizes a transformer-based model—specifically GPT-2—fine-tuned for sequence classification. Figure 3 illustrates that the input text undergoes initial preprocessing, followed by tokenization via the GPT-2 tokenizer, which use the end-of-sequence token for padding and limits sequences to 512 tokens. A data collator dynamically manages padding at runtime. The generated token IDs and attention masks are subsequently fed into GPT2ForSequenceClassification, a model that incorporates a classification head on top of the pre-trained GPT-2 transformer. Leveraging the extensive contextual embeddings acquired from unsupervised pre-training, the model may be fine-tuned to effectively differentiate between AI-generated text and human-written information.

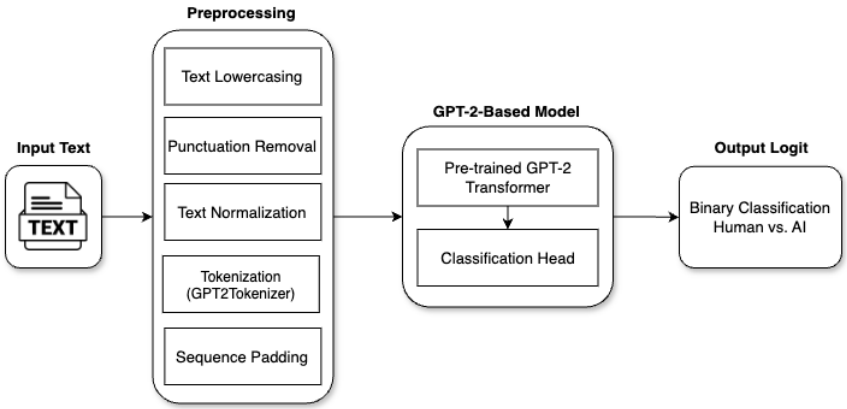

Figure 3 – GPT-2-based detection model

Experimental Setup

All models were implemented via PyTorch and trained on GPU-enabled hardware to guarantee efficient computation. The training procedure was the same across models: each underwent training for 10 epochs with consistent data preprocessing, normalization, tokenization, and sequence padding to ensure a fair comparison. The Adam optimizer, with a learning rate of 1e-3, was used for the LSTM-based and CNN-LSTM hybrid models, while training stability was further enhanced by gradient clipping (maximum norm of 1.0) and a ReduceLROnPlateau learning rate scheduler. The GPT-2 model was fine-tuned with the Hugging Face Trainer API with a learning rate of 2e-5 and batch sizes of 4 for training and 8 for assessment. Evaluation measures, such as accuracy, precision, recall, and F1-score, were calculated on the validation set utilizing standard functions from scikit-learn, while hyperparameters were refined by grid search to guarantee optimal performance.

To systematically assess model performance, we utilize four established assessment metrics—Accuracy, Precision, Recall, and F1-score—as delineated below:

Accuracy reflects the overall correctness of classification and is computed as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

where TP and TN are true positives and true negatives, and FP and FN are false positives and false negatives, respectively.

Precision indicates the proportion of texts labeled as AI-generated that truly are AI-generated:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

with TP as correctly identified AI-generated texts and FP as misclassified human texts.

Recall (sensitivity) measures the fraction of actual AI-generated texts detected:

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

where FN are AI-generated texts missed by the model.

F1-score harmonizes Precision and Recall into a single value:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

balancing both false positivs and false negatives.

Results and their discussion.

We evaluated the performance of the three models—GPT-2-based, LSTM-based, and CNN-LSTM hybrid—on identifying AI-generated text using the standardized dataset outlined in Section III.A. The research concentrated on quantifying essential performance parameters, including accuracy, precision, recall, and F1-score on the validation set. Table 2 encapsulates the performance metrics: The GPT-2-based model attained an accuracy of 96.88%, with precision, recall, and F1-score of 96.96%, 96.74%, and 96.85%, respectively; the CNN-LSTM hybrid model achieved an accuracy of 95.16% and an F1-score of 94.94%; whereas the LSTM-based model secured an accuracy of 94.64% and an F1-score of 94.29%.

Table 2 – Performance Comparison of GPT - 2, CNN - LSTM, and LSTM Models

Model	Accuracy	Precision	Recall	F1-Score
GPT - 2 Based	96.88%	96.96%	96.74%	96.85%
CNN - LSTM Hybrid	95.16%	94.28%	96.61%	94.94%
LSTM - Based	94.64%	93.43%	95.17%	94.29%

Figure 4 displays a bar chart that contrasts the performance metrics—Accuracy, Precision, Recall, and F1-Score—of the three models: GPT-2-based, CNN-LSTM hybrid, and LSTM-based. The chart distinctly illustrates that the GPT-2-based model maintains an overall performance advantage across all metrics, while the CNN-LSTM hybrid and LSTM-based models follow closely with slightly lower values.

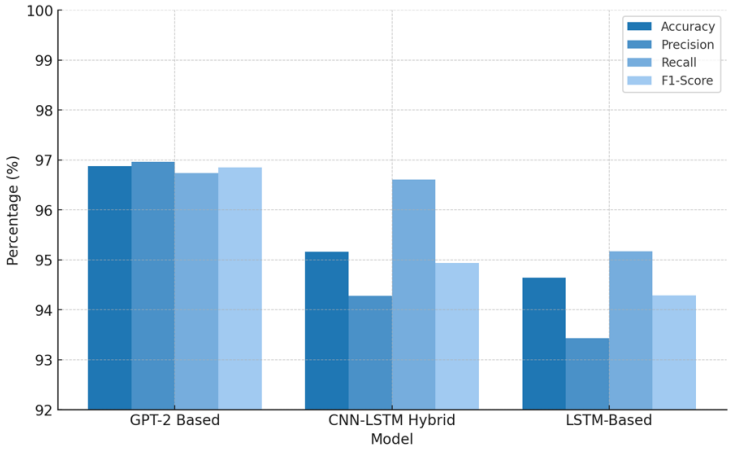

Figure 4 – GPT-2, CNN-LSTM, and LSTM Model Performance Metrics

A representation of the training loss curves for each model over the course of ten epochs is presented in Figure 5. The line graph illustrates the convergence behavior and stability of the training process, with all models exhibiting a consistent reduction in loss as training advances.

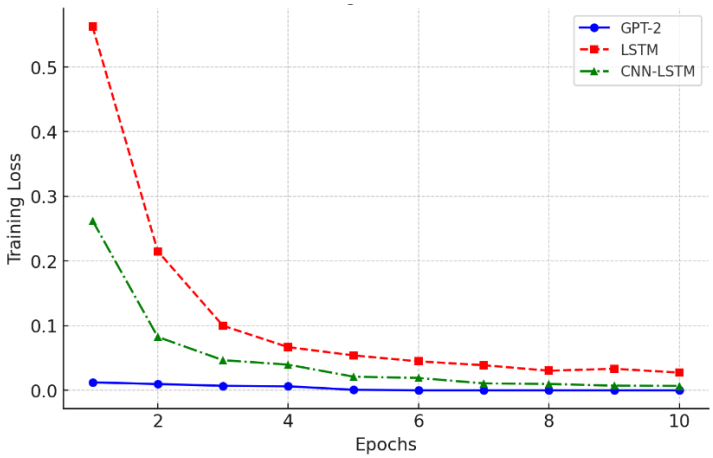

Figure 5 – Training Loss Curves for GPT-2, LSTM, and CNN-LSTM Models

As seen in Figure 6, the F1 score for each of the three models has changed during the course of the training epochs. This graph demonstrates the enhancements in classification performance throughout fine-tuning, further validating the robustness of each approach.

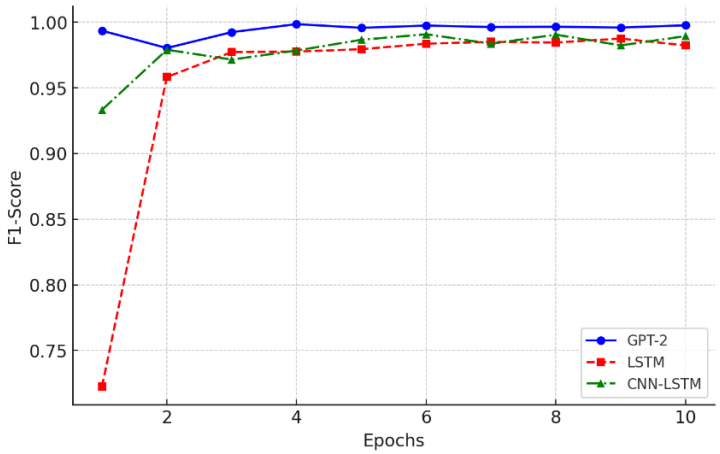

Figure 6 – F1-Score Progression Over Training Epochs for GPT-2, LSTM, and CNN-LSTM Models

To clarify the individual failure scenarios of each model, we provide a comprehensive error analysis in Figure 7. The GPT-2 detector demonstrates minimal misclassifications, recording only one false positive and six false negatives, suggesting that its deep contextual representations proficiently disambiguate the majority of samples, however it may sometimes misread exceedingly concise human-written or AI-generated texts. The CNN-LSTM hybrid has a greater yet balanced error profile, including thirty false positives and thirteen false negatives, illustrating the trade-off associated with merging local feature extraction with sequential modeling. The LSTM baseline, although computationally efficient, has the highest total mistake count (sixty false positives and nineteen false negatives), indicating challenges in recognizing the subtle textual patterns that differentiate human and AI outputs. Figure 7 quantifies different mistake kinds, emphasizing specific goal areas: minimizing false positives for GPT-2, adjusting threshold sensitivity for the hybrid model, and enhancing context modeling for LSTM, thereby directing future improvements in detection robustness.

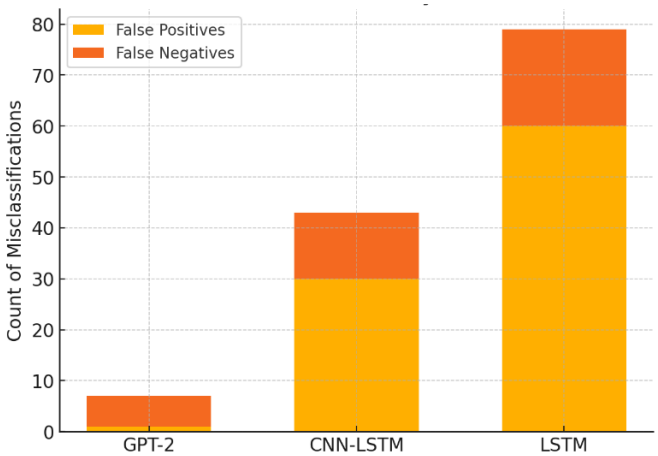

Figure 7 – Misclassification Breakdown for AI-Text Detection Models

Clear trade-offs between the three approaches are shown by a comparative analysis. Although the GPT-2-based model requires more computer resources and longer training times, it uses deep contextual representations acquired by transfer learning to produce somewhat improved results. By combining local feature extraction with sequential modeling, the CNN-LSTM hybrid model offers a good balance between computational economy and excellent accuracy. On the other hand, even if it is slightly less accurate, the LSTM-based model provides a simpler and more

economical baseline The results show that specific application demands, including the requirement for high accuracy compared to available computer resources, should guide model selection.

In order to evaluate our detectors' resilience in a real-world educational setting, we put them to the test using twenty anonymized student assignments from an undergraduate course and concurrently produced twenty comparable AI-generated works for the same assignment with a GPT-based text generator. Upon assessment using both the CNN-LSTM and GPT-2 detectors, authentic student submissions exhibited AI-generation probabilities ranging from 72 % to 86 %, indicative of occasional templated phrases, whereas the AI-generated submissions for identical assignments attained scores of 96 % to 98 %, thereby validating their artificial origin. Moving forward, incorporating additional data from diverse disciplines into our training set will further improve model calibration, reduce false positives, and adapt to new writing styles and assignment formats.

It is crucial to balance the computational demands and the infrastructure that is available when implementing these models in real-world learning environments. Transformer-based classifiers impose considerable memory and processing demands due to their multi-headed attention methods and deep architecture, occasionally requiring specialized accelerators or scalable cloud resources to provide satisfactory response times. In contrast, a hybrid architecture that integrates convolutional feature extractors with a recurrent layer can function effectively on less powerful hardware—such as ordinary servers or CPU-only systems—without significantly compromising detection performance. The hybrid method is especially appropriate for schools that need to reconcile academic integrity measures with financial and operational constraints. Model selection should ultimately be informed by an institution's available computational resources, maintenance capabilities, and the required level of detection robustness.

Considering the trade-offs in precision, computational demands, and intricacy, we utilized GPT-2 for its superior detection capabilities, the CNN-LSTM hybrid for its ideal equilibrium of efficiency and resilience, and the LSTM model for its simplicity and low resource requirements - together accommodating a broad spectrum of institutional needs.

Conclusion.

Finally, our comparison of three deep learning approaches—GPT-2-based, CNN-LSTM hybrid, and LSTM-based models—for recognizing AI-generated text shows that every strategy has better performance on the validation set. Thanks to its powerful contextual representations and transfer learning capacity, the GPT-2-based model attained the highest accuracy and F1-score. Concurrently, via clever integration of local feature extraction and sequential modeling, the CNN-LSTM hybrid model obtained a great balance between accuracy and computing efficiency. Especially in environments with limited computing resources, the LSTM-based model offers a dependable baseline even if its relative simplicity and minimal resource needs show a lower raw accuracy.

Especially in higher education, these results have major implications for artificial intelligence text detection. Maintaining academic integrity in digital learning environments depends on the ability to autonomously and accurately distinguish between human-generated and artificial intelligence-generated material. Our results show that although they somewhat increase performance, transformer-based models require more computing resources. On practical uses where resource constraints are major, hybrid systems are appealing solutions as they can provide equivalent accuracy with less complexity.

Future research will aim to extend the evaluation scope by incorporating bigger and more diverse datasets, exploring other hybrid model configurations, and refining hyperparameter tuning to enhance detection accuracy and adaptability in real educational settings.

References

1. Mishra, S. (2023). Enhancing Plagiarism Detection: The Role of Artificial Intelligence in Upholding Academic Integrity. *Library Philosophy & Practice*.

2. Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*(1), 26. https://doi.org/10.1007/s40979-023-00146-z.

3. Luan, Y., & Lin, S. (2019). Research on text classification based on CNN and LSTM. *In Proceedings of the 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)* (pp. 352–355). IEEE. https://doi.org/10.1109/ICAICA.2019.8873454.

4. Alshingiti, Z., Alaqel, R., Al-Muhtadi, J., Haq, Q. E. U., Saleem, K., & Faheem, M. H. (2023). A Deep Learning-Based Phishing Detection System Using CNN, LSTM, and LSTM-CNN. *Electronics*, 12(1), 232. https://doi.org/10.3390/electronics12010232.

5. Tang, R., Chuang, Y.-N., & Hu, X. (2024). The science of detecting LLM-generated text. *Communications of the ACM*, 67(4), 50–59. https://doi.org/10.1145/3624725.

6. Zulqarnain, M., Ghazali, R., Hassim, Y. M. M., & Rehan, M. (2020). A comparative review on deep learning models for text classification. *Indonesian Journal of Electrical Engineering and Computer Science*, 19(1), 325–335. https://doi.org/10.11591/ijeecs.v19.i1.pp325-335

7. Kayabas, A., Topcu, A. E., Alzoubi, Y. I., & Yıldız, M. (2025). A Deep Learning Approach to Classify AI-Generated and Human-Written Texts. *Applied Sciences*, 15(10), 5541. https://doi.org/10.3390/app15105541.

8. Utku, A. (2024). Hybrid CNN-LSTM model for fake news detection. *Malatya Turgut Özal University Journal of Engineering and Natural Sciences*, 5(2), 28–36. https://doi.org/10.46572/naturengs.1571897.

9. Walker, E., Evans, L., Mitchell, A., Zhang, Z., Patel, R., & Chen, I. (2024). Text classification in detection of AI-generated content using BERT [Manuscript under review]. *Machine Learning and Systems Conference*. https://doi.org/10.13140/RG.2.2.30193.90722.

10. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*

11. Guerrero, J. A. (2023). Detecting AI generated text using neural networks (Master's thesis, Texas A&M University).

12. Kaibassova, D., & Nurtay, M. (2022). The comparative analysis of machine learning models for quality assessment of textual academic works. *In Proceedings of the 2022 International Conference on Smart Information Systems and Technologies (SIST)* (pp. 1–4). IEEE. https://doi.org/10.1109/SIST54437.2022.9945714.

13. Mulenga, R., & Shilongo, H. (2024). Academic integrity in higher education: Understanding and addressing plagiarism. *Acta Pedagogia Asiana*, 3(1), 30–43. https://doi.org/10.53623/apga.v3i1.337

14. Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education*, 20, Article 38. https://doi.org/10.1186/s41239-023-00408-3.

15. Temper, M., Tjoa, S., & David, L. (2025). Higher Education Act for AI (HEAT-AI): A framework to regulate the usage of AI in higher education institutions. *Frontiers in Education*, 10, Article 1505370. https://doi.org/10.3389/feduc.2025.1505370.

16. Thite, S. (2023). LLM – Detect AI generated text dataset (Version 1) [Data set]. Kaggle. https://www.kaggle.com/datasets/sunilthite/llm-detect-ai-generated-text-dataset/data.

СТУДЕНТТЕРДІҢ МӘТІНДІК ТАПСЫРМАЛАРЫНДАҒЫ ЖАСАНДЫ ИНТЕЛЛЕКТІМЕН ЖАСАЛҒАН МӘТІНДІ АНЫҚТАУ МОДЕЛЬДЕРІНІҢ САЛЫСТЫРМАЛЫ ТАЛДАУЫ

Аңдатпа. *Қазіргі трансформерлік модельдер автоматты мәтін генерациясының мүмкіндіктерін едәуір кеңейтіп, жоғарғы оқу орындарында академиялық адалдықты*

қамтамасыз етуде жаңа сынақтарды туындатты. Дәстүрлі антиплагиат жүйелері студенттің өзіндік жұмысы мен жасанды интеллектпен жасалған материалды ажырата алмайды, сондықтан сенімді автоматты анықтаушылар әзірлеу өзекті болып табылады. Осыған байланысты мақалада студенттік жұмыстардағы ИИ-генерацияланған мәтінді анықтаудың үш негізгі әдісінің салыстырмалы талдауы жүргізілді. Зерттеу нысаны – GPT-2 трансформеріне негізделген алдын ала оқытылған классификатор, CNN-LSTM гибридтік архитектура және классикалық LSTM-моделі. Зерттеу міндеттері: эксперименттік протоколды біріздендіру, деректерді сапалы алдын ала өңдеу және әр тәсілдің есептеу қуаты мен дәлдік талаптарына сәйкес тиімділігін жан-жақты бағалау. Зерттеу әдістемесі алдын ала өңделген студенттік жұмыстар корпусын бірегей токенизация және оптимизация параметрлерімен оқу және тексеру жиынтықтарына бөлу, модельдерді бірнеше эпохада бірдей гиперпараметрлермен дайындау, нәтижелерді дәлдік, қамту және F1-көрсеткіштері бойынша салыстырудан тұрады. Нәтижелер көрсеткендей, трансформерлік детектор ең терең контекстік бейнелеуді қамтамасыз етіп, ең жоғары дәлдік көрсетіп, CNN-LSTM гибриді жылдамдық пен сапаны тиімді теңестірсе, LSTM-моделі GPU-қолдаусыз ортада ресурсты үнемдейтін базалық шешім ретінде оңтайлы екенін дәлелдеді. Қорытындысында авторлар инфрақұрылым талаптарына сәйкес әдісті таңдау маңызды екенін атап, жоғары өнімді GPU серверлері үшін трансформерлік архитектураларды, орта деңгейлі аппараттық платформада гибридтік шешімдерді, ал тек CPU-негізіндегі ортада LSTM-модульдерді қолдануды ұсынады. Практикалық ұсыныс ретінде гибридтік детекторды білім беру платформаларына сараптамалық рецензиямен бірге енгізіп, оқыту дерекқорын үнемі жаңартып отыруды ұсынады.

***Түйін сөздер**: Терең оқыту, Жасанды интеллектпен жасалған мәтін, Академиялық адалдық, GPT-2, CNN-LSTM гибридтік моделі, LSTM, Мәтінді жіктеу, Трансформерлік модельдер, Гибридтік модельдер.*

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МОДЕЛЕЙ ОБНАРУЖЕНИЯ ТЕКСТА, СГЕНЕРИРОВАННЫХ ИИ, В ТЕКСТОВЫХ ЗАДАНИЯХ СТУДЕНТОВ

***Аннотация.** Современные трансформерные модели существенно расширили возможности автоматической генерации текстов, что создаёт новые вызовы для обеспечения академической честности в вузах. Системы антиплагиата нередко не различают работы, написанные студентом, и материалы, сгенерированные искусственным интеллектом, что обуславливает актуальность разработки надёжных автоматических детекторов. В связи с этим в данной статье выполнен сравнительный анализ трёх подходов к обнаружению ИИ-генерированного текста в студенческих работах. Предмет исследования — особенности работы классификатора на базе GPT-2, гибридной архитектуры CNN-LSTM и классической LSTM-модели. Задачи включают формирование единого экспериментального протокола и оценку каждого метода в условиях ограниченных вычислительных ресурсов и различных требований к точности. Экспериментальная методика исследовательской работы состоит из единообразной предобработки размеченного корпуса студенческих работ, разделения данных на обучающую и валидационную выборки, обучения моделей в несколько эпох с одинаковыми параметрами токенизации и оптимизации, а также оценки их эффективности по показателям точности, полноты и F1-меры. Результаты исследования показывают, что детектор на основе трансформера обеспечивает наиболее глубокое контекстное представление, гибридный CNN-LSTM демонстрирует оптимальный баланс между скоростью обработки и качеством обнаружения, а LSTM-модель остаётся эффективным и ресурсосберегающим решением для систем без доступа к GPU. Авторы пришли к выводу, что выбор метода должен основываться на доступной инфраструктуре: трансформеры*

подходят для высокопроизводительных серверов с GPU, гибридные архитектуры — для платформ средней производительности, а LSTM-модули — для CPU-окружения. В качестве практической рекомендации авторами предлагается интегрировать гибридный детектор в образовательные платформы совместно с экспертным рецензированием и регулярно обновлять обучающую базу для адаптации к новым типам ИИ-контента.

Ключевые слова: *Глубокое обучение, Текст, сгенерированный ИИ, Академическая честность, GPT-2, Гибридная модель CNN-LSTM, LSTM, Классификация текста, Трансформерные модели, Гибридные модели.*

Сведение об авторах

Сагынбаев Алмаз Асылбекулы	Магистрант, Департамент Компьютерной Инженерии, Astana IT университет, Астана, Казахстан, E-mail: 232040@astanait.edu.kz
Кайбасова Динара Женисбековна	PhD, Доцент, Департамент Компьютерной инженерии, Astana IT университет, Астана, Казахстан, E-mail: dinara.kaibasova@astanait.edu.kz
Сагынбаева Айтолкын Асылбеккызы	Сеньор-лектор, Школа Креативной Индустрии, Астана IT университет, Астана, Казахстан, E-mail: Aitolkyn.sagynbayeva@astanait.edu.kz

Авторлар туралы мәлімет

Сағынбаев Алмаз Асылбекұлы	Магистрант, Компьютерлік Инженерия Департаменті, Astana IT университеті, Астана, Қазақстан,. E-mail: 232040@astanait.edu.kz
Кайбасова Динара Женисбековна	PhD, Компьютерлік Инженерия Департаменті, Astana IT университетінің қауымдастырылған профессоры, Астана, Қазақстан E-mail: dinara.kaibasova@astanait.edu.kz
Сағынбаева Айтолкын Асылбекқызы	Сеньор-лектор, Шығармашылық Индустрия Мектебі, Астана IT университеті, Астана, Қазақстан, E-mail: Aitolkyn.sagynbayeva@astanait.edu.kz

Information about the authors

Sagynbayev Almaz Assylbekuly	Master's student, Department of Computer Engineering, Astana IT University, Astana, Kazakstan, E-mail: 232040@astanait.edu.kz
Kaibassova Dinara Zhenisbekovna	PhD, Associate Professor, Department of Computer Engineering in Astana IT University, Astana, Kazakstan, E-mail: dinara.kaibasova@astanait.edu.kz
Sagynbayeva Aitolkyn Assylbekkyzy	Senior lecturer, School of Creative Industries, Astana IT University, Astana, Kazakstan E-mail: Aitolkyn.sagynbayeva@astanait.edu.kz