2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST)
14-16 May, 2025, Astana, Kazakhstan

Text Classification for AI Generated Content with Machine Learning and Deep Learning Models

Batyr Sharimbayev
Department of Mathematics and Natural Sciences
SDU University
Kaskelen, Kazakhstan
batyr.sharimbayev@sdu.edu.kz
0009-0006-3323-231X

Shirali Kadyrov
Department of General Education
New Uzbekistan University
Tashkent, Uzbekistan
sh.kadyrov@newuu.uz
0000-0002-8352-2597

***Abstract*—The rapid development of generative AI models, such as GPT-4, LLaMA, and Gemini, is causing an explosion of AI-generated text that may be akin to human writing. This poses a challenge in differentiating between AI-generated content and human-authored text across a range of verticals: academic integrity, misinformation detection, and content moderation. This paper presents a comparison of machine learning and deep learning models on the classifier for AI-generated text. We compare the performance of Logistic Regression with TF-IDF features, a Bi-LSTM model, and a fine-tuned DistilBERT model on data from the COLING Workshop on MGT Detection Task 1, involving text samples from five AI models and human authors. Our experiments showed that Bi-LSTM outperforms other models, yielding the best results in accuracy (90.09%) and F1-score (90.02%). We further present the binary classification performance that distinguishes AI-generated text from human-written content, with an accuracy of 95.9%. It is suggested that deep learning methods are competent in detecting AI-generated text, though there are certain limitations, including adversarial attacks and changing styles of AI-generated writing. Future work will be focused on enhancing model robustness through adversarial training and hybrid architectures.**

Keywords—AI-generated text, text classification, deep learning, Bi-LSTM, DistilBERT, machine learning

I. INTRODUCTION

With the advent of generative AI models like LLaMA [1], Grok (xAI), Gemini (Google DeepMind) [2], GPT-4 (OpenAI), and DeepSeek-V3 [3], the amount of AI-generated literature that closely mimics human writing has greatly grown. Strong detection techniques are required because of the difficulties in differentiating between human and machine-generated content caused by the quick development of generative AI [4, 5]. In fields like content moderation, academic integrity, and disinformation prevention, accurate identification of AI-generated material is essential. Deep learning (DL) and machine learning (ML) approaches have been extensively investigated for text categorization tasks in order to solve this problem, allowing for the identification of AI-generated content with differing levels of accuracy.

Early efforts in text classification relied on statistical and rule-based methods, such as Support Vector Machines (SVM) [6] and Naïve Bayes [7]. These approaches typically leverage handcrafted features such as word frequency, n-grams, and linguistic markers to distinguish different types of text. While effective for simpler classification tasks, these models often struggle with more nuanced patterns present in AI-generated content [8].

Feature engineering has played a crucial role in enhancing traditional models. Mindner et al. [9] achieved high F1 scores using a combination of perplexity, semantic, and readability features, among others. Shah et al. [10] explored syllable count, word length, and functional word usage, achieving 93% accuracy. Liu & Kong [11] employed a strided sliding window approach to extract perplexity features, effectively distinguishing between human and AI-generated content. However, with the emergence of more advanced AI models, these handcrafted features are becoming less effective, necessitating the adoption of deep learning-based methods.

The performance of neural networks in text classification has greatly improved, especially transformer-based models. To improve detection accuracy, BERT [12], RoBERTa [13], and XLNet [14] have been refined on datasets of human-written and AI-generated text. For instance, Javaji et al. [15] concentrate on creating a cutting-edge machine learning model based on BERT to differentiate between articles written by students and those produced by large language models (LLMs).

More recent work has leveraged advanced deep learning architectures, including CNNs, LSTMs, GRUs, and BiLSTMs, along with various text embedding techniques to detect AI-generated content. The BiLSTM model achieved the best performance [16]. However, the reliability of AI text detectors has been questioned due to

evasion techniques such as recursive paraphrasing attacks, intentional errors to mimic human writing imperfections, adjustments in text complexity to resemble human styles, and the use of automated paraphrasing tools to significantly alter content [17].

The evolving nature of generative models necessitates the development of dynamic and continuously trained classification techniques [18]. Despite achieving high accuracy in some experiments, classifiers still struggle with certain datasets, highlighting the ongoing challenges in this field [19]. As AI-generated content becomes more prevalent, addressing security, privacy, ethical, and legal concerns becomes crucial, with watermarking approaches emerging as potential solutions for regulatable AIGC paradigms [20].

Our goal in this study is to assess how well ML and DL classifiers perform when classifying a given text into five different classes: human-written text and four cutting-edge AI text generators: GPT-3.5-turbo, Bloomz, Cohere, and Davinci. We may examine in greater detail how well various models will be able to differentiate between particular AI-generated material and a text written by a human thanks to this multi-class classification. As a secondary objective, we also explore the binary classification problem, which reflects a simpler but highly practical task: identifying any given text as AI-generated versus human-written. This binary classification is of greater relevance for a number of practical applications, such as plagiarism detection, content moderation, and misinformation prevention, where, in general, stakeholders are just interested in telling whether a certain text was generated by AI, not by exactly which generator.

Specifically, we will focus on determining the effective methodology for detecting AI-generated text through evaluating and comparing various traditional ML models with state-of-the-art DL architectures. Our analysis will enrich the automatic improvement of detection frameworks with useful insights for scholars, educators, and organizations in their quest to preserve content authenticity and integrity amidst the upsurge of AI-sourced text. The task of accurately detecting AI-generated text presents considerable challenges, especially due to the rise of techniques such as text humanization tools that are designed to obscure machine-generated patterns and make the output resemble human writing. This complicates the binary classification of authorship between humans and AI systems. We propose that an effective strategy to mitigate this issue is to first identify the specific generative model responsible for the text. By doing this, the classifier can pick up linguistic and stylistic traits unique to the model, improving its capacity to discern between writing produced by AI and text written by humans.

II. DATA PREPARATION

A. Data Collection

The dataset from Subtask A of the COLING Workshop on MGT Detection Task 1 was used for English-only machine-generated text detection [21]. However, only the five models with the most data were classified, resulting in a dataset containing a total of 204,408 text samples. Of these, gpt-3.5-turbo had 59,534, human had 57,563, bloomz had 30,052, cohere had 29,704, and davinci had 27,555 samples. The dataset does not contain missing values and is stored in JSON format.

B. Data Preprocessing

Different preprocessing steps were carried out on the dataset, aimed at cleansing and normalizing text data into a workable form. It started with cleaning of text-removal of non-alphanumeric characters, removal of digits, and removal of extra changing cases to lower for uniformity. Then, remove stopwords with the help of NLTK which removes irrelevant content. Lemmatization-that process whereby the summary of the word in its base form will be done; for example, "running" was reduced to "run.

The average text length is 1,115 characters, with a median of 905. Texts range from 0 to 13,104 characters, with most falling between 589 and 1,424 characters. This distribution, as shown in Figure 1.

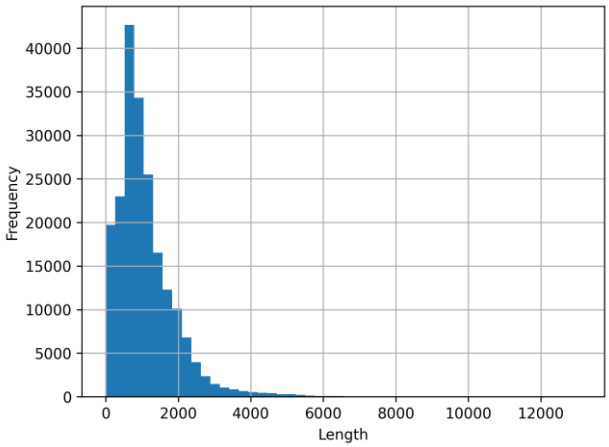

Figure 1. Distribution of text lengths of the dataset.

III. MODEL SELECTION

They are different natures of ML classifiers and DL models, hence the basis for model selection. At first, Logistic Regression with TF-IDF [22] features based on word n-grams was considered to appraise its performance in text classification. It weighs term frequency by inverse document frequency, an adjusting factor that diminishes terms which occur in large numbers of documents in an entire collection of them. Because of this, TF-IDF is effective for text classification and information retrieval. The method will help in highlighting key terms that contribute most to distinguishing between AI-generated and human-written text.

Subsequently, more advanced models, including the Bi-LSTM (Bidirectional Long Short-Term Memory) and the pre-trained DistilBERT model, were explored to capture complex text patterns. Bi-LSTM allows contextual understanding by processing text, while DistilBERT, a lighter version of BERT, leverages self-attention mechanisms to improve text representation [23]. In the case of DistilBERT, fine-tuning was done only on the last two layers to balance model performance and computational efficiency. Further, both models were evaluated based on accuracy and macro/micro F1-scores, which are crucial for ensuring a robust comparison across different classification tasks.

The Logistic Regression model is trained on text data, which was transformed by the TF-IDF vectorization.

Feature extraction from text includes 1-grams and 2-grams, limited to 5000 features to ensure computational efficiency.

TABLE I. BI-LSTM MODEL ARCHITECTURE

Layer	Output shape	Params #
Embedding	(None, 512, 64)	320,000
Bi-LSTM	(None, 512, 128)	66,048
Bi-LSTM	(None, 128)	98,816
Dropout	(None, 128)	0
Dense	(None, 32)	4,128
Dense	(None, 5)	165

A Bi-LSTM, which is well-suited for classification problems requiring sequence data, is used in the model architecture for the deep learning technique [24, 25]. To guarantee consistent input shape, sequences are first padded to a fixed length of 512 tokens after the data has been tokenized with a maximum vocabulary size of 5000 words. An embedding layer, two bidirectional LSTM layers, dropout layers to lessen overfitting, and a final dense layer are all included in the model. For multi-GPU training, it employs MirroredStrategy and is compiled with the Adam optimizer with Sparse Categorical Crossentropy loss. As shown in Table I, the model comprises 489,157 trainable parameters and divides text into five groups.

The final model employs DistilBERT, a smaller version of BERT, for text classification. The model uses a sequence length of 512 tokens, with a batch size of 32, and is trained for 3 epochs with a learning rate of 0.0005. A DistilBERT Preprocessor is applied to tokenize the text data, followed by a DistilBERT Classifier for multi-class classification. To fine-tune the model, only the last two layers of the DistilBERT backbone are made trainable, while the other layers remain frozen. The model is compiled with Sparse Categorical Crossentropy loss and Adam optimizer.

IV. RESULTS

Our study's findings offer a thorough assessment of how well both conventional ML and DL models identify AI-generated text.. Our findings highlight key differences in detection capabilities across models, revealing which approaches are more adept at distinguishing between human-written text and outputs from AI systems such as GPT-3.5-turbo, Bloomz, Cohere, and Davinci. Additionally, we analyze error patterns and misclassification trends, shedding light on the challenges associated with AI text detection and the implications for real-world applications.

Table II shows the performance comparison of the three models. Bi-LSTM outperformed both DistilBERT and Logistic Regression with TF-IDF, achieving the highest scores across all metrics. DistilBERT performed well, benefiting from contextual embeddings, while Logistic Regression with TF-IDF had the lowest scores.

TABLE II. PERFORMANCE OF THE MODELS

Model	Macro-F1	Micro-F1	Accuracy
LogReg + TF-IDF	80.94%	81.45%	81.45%
DistilBERT	86.76%	87.61%	87.61%
Bi-LSTM	**90.02%**	**90.09%**	**90.09%**

Five epochs were used to train the Bi-LSTM model. The model's validation accuracy reached 90.10%, while its training accuracy rose from 58.16% in the first epoch to 92.43% in the last. the consistent improvement in accuracy and the reduction in validation loss.

Figure 2 presents the confusion matrix for the Bi-LSTM model, showing how well it classifies different labels. The diagonal values represent correct predictions, with high counts for each class, indicating strong performance. Misclassifications are relatively low but occur mainly between similar categories, such as bloomz and gpt. The model performs best at distinguishing "human" text, with 10,871 correct predictions, while some overlap exists among AI-generated texts.

Figure 3 illustrates the classification accuracy for different categories. The Bi-LSTM model achieved the highest accuracy 95.70% on bloomz, suggesting that its textual patterns were the most distinguishable.

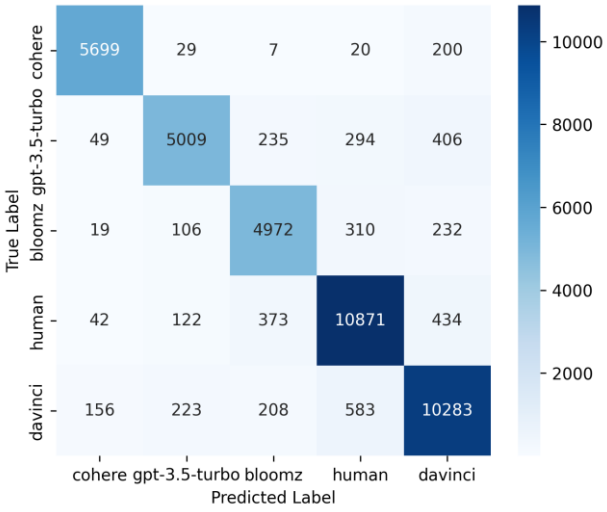

Figure 2. Confusion matrix for Bi-LSTM model

Figure 3 shows the classification accuracy for different categories. The best performance was achieved for bloomz (95.70%), indicating it was the easiest to classify. The worst performance was for cohere (83.58%)

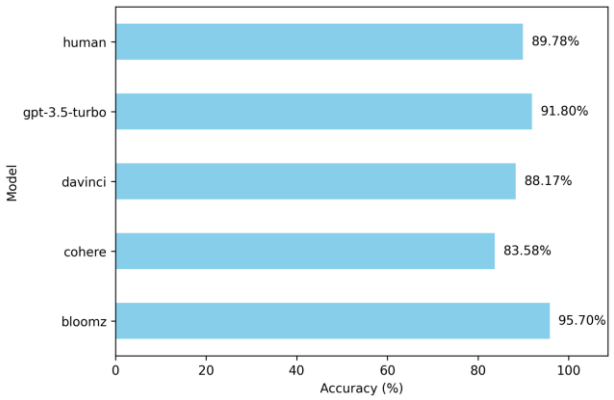

Figure 3. Accuracy of correctly predicted models

Consequently, we also present the Bi-LSTM model's binary classification results in separating text produced by AI from text authored by humans in Table III.

TABLE III. PERFORMANCE OF THE BI-LSTM MODEL FOR BINARY TEXT CLASSIFICATION

Model	Macro-F1	Micro-F1	Accuracy
Bi-LSTM	92.7%	95.5%	95.9%

The model's performance in distinguishing human-written from AI-generated text was assessed using a 2×2 confusion matrix. As shown in Figure 4, the model correctly classified 10,871 human texts and 25,963 AI texts, while 1,933 AI texts were misclassified as human, and 537 human texts were mistaken for AI.

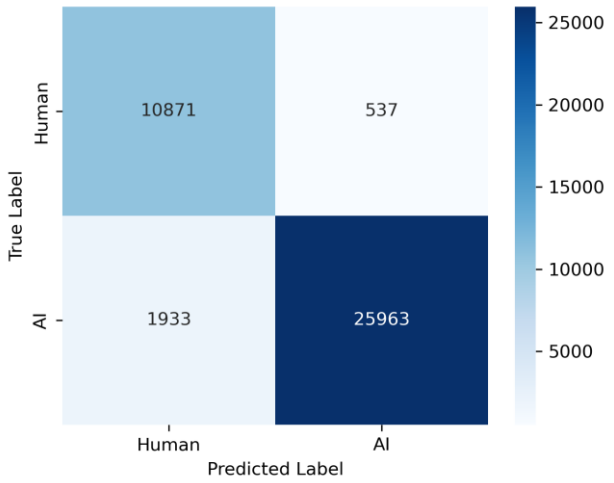

Figure 4. 2×2 Confusion matrix for human vs AI classification

V. CONCLUSION

This research presented a comparative analysis of ML and DL models regarding classifying AI-generated text. We tested the performance of Logistic Regression with TF-IDF, Bi-LSTM, and DistilBERT on a dataset containing texts from GPT-3.5-turbo, human, Bloomz, Cohere, and Davinci models. Our results show that deep learning approaches, and more precisely Bi-LSTM, significantly outperform classic machine learning methods when it comes to the detection of AI-generated text. The Bi-LSTM model achieved the highest accuracy and F1 scores, effectively capturing contextual dependencies in text classification tasks.

Despite these positive results, AI-generated text detection still has various challenges. As generative models continuously evolve, classifiers must adapt to increasingly sophisticated text patterns that begin to resemble human writing. An adversarial training process with continuous learning and hybrid models incorporating linguistic and statistical features may improve their robustness against evasion methods. Moreover, AI methodologies can also help to make models explainable by revealing insight into the decision-making process

Future research should focus on multi-modal approaches that can integrate metadata, stylometric analysis, and ensemble learning for greater improvement in the performance of detection. Investigating the cross-lingual and domain-specific detection capabilities may help raise generalizability. Importantly, fairness, transparency, and regulatory compliance are ethical considerations to which attention needs to be paid for the responsible deployment of any AI detection system. These will lead to greater reliability in adaptive solutions that can make a distinction between AI-generated content and human-authored text.

ACKNOWLEDGEMENT

This research work is supported by project AP23487777 of the Ministry of Science and Higher Education of the Republic of Kazakhstan

REFERENCES

[1] H. Touvron et al., “LLaMA: Open and Efficient Foundation Language Models,” arXiv:2302.13971 [cs], Feb. 2023, Available: https://arxiv.org/abs/2302.13971.

[2] Gemini Team et al., “Gemini: A Family of Highly Capable Multimodal Models,” arXiv.org, Dec. 18, 2023. https://arxiv.org/abs/2312.11805.

[3] DeepSeek-AI et al., “DeepSeek LLM: Scaling Open-Source Language Models with Longtermism,” arXiv (Cornell University), Jan. 2024, doi: https://doi.org/10.48550/arxiv.2401.02954.

[4] H. Hua and C.-J. Yao, “Investigating generative AI models and detection techniques: impacts of tokenization and dataset size on identification of AI-generated text,” Frontiers in Artificial Intelligence, vol. 7, Nov. 2024, doi: https://doi.org/10.3389/frai.2024.1469197.

[5] F. F.-H. Nah, R. Zheng, J. Cai, K. Siau, and L. Chen, “Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration,” Journal of information technology case and application research, vol. 25, no. 3, pp. 277–304, Jul. 2023, doi: https://doi.org/10.1080/15228053.2023.2233814.

[6] T. Joachims, “Text categorization with Support Vector Machines: Learning with many relevant features,” Machine Learning: ECML-98, vol. 1398, pp. 137–142, 1998, doi: https://doi.org/10.1007/bfb0026683.

[7] A. Mccallum and K. Nigam, “A comparison of event models for naive bayes text classification,” 1998.

[8] B. Tom et al., “Language Models are Few-Shot Learners,” Advances in Neural Information Processing Systems, vol. 33, 2020.

[9] Lorenz Mindner, T. Schlippe, and K. Schaaff, “Classification of Human- and AI-Generated Texts: Investigating Features for ChatGPT,” Lecture notes on data engineering and communications technologies, pp. 152–170, Jan. 2023, doi: https://doi.org/10.1007/978-981-99-7947-9_12.

[10] A. Shah et al., “Detecting and unmasking AI-generated texts through explainable artificial intelligence using stylistic features,” Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 10, 2023.

[11] X. Liu and L. Kong, “AI Text Detection Method Based on Perplexity Features with Strided Sliding Window Notebook for the PAN Lab at CLEF 2024.
[12] M. Khadhraoui, H. Bellaaj, M. B. Ammar, H. Hamam, and M. Jmaiel, “Survey of BERT-Base Models for Scientific Text Classification: COVID-19 Case Study,” Applied Sciences, vol. 12, no. 6, p. 2891, Jan. 2022, doi: https://doi.org/10.3390/app12062891.
[13] J. Briskilal and C. N. Subalalitha, “An ensemble model for classifying idioms and literal texts using BERT and RoBERTa,” Information Processing & Management, vol. 59, no. 1, p. 102756, Jan. 2022, doi: https://doi.org/10.1016/j.ipm.2021.102756.
[14] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, “XLNet: Generalized Autoregressive Pretraining for Language Understanding,” Advances in Neural Information Processing Systems, vol. 32, 2019.
[15] P. Javaji, P. S. Sreeya, and S. Rajesh, “Detection of AI Generated Text With BERT Model,” 2024 2nd World Conference on Communication & Computing (WCONF), pp. 1–6, Jul. 2024, doi: https://doi.org/10.1109/wconf61366.2024.10692072.
[16] M. A. Wani, M. ElAffendi, and K. A. Shakil, “AI-generated spam review detection framework with deep learning algorithms and natural language processing,” Comput., vol. 13, p. 264, 2024.
[17] M. Perkins et al., “Simple techniques to bypass GenAI text detectors: implications for inclusive education,” International Journal of Educational Technology in Higher Education, vol. 21, no. 1, Sep. 2024, doi: https://doi.org/10.1186/s41239-024-00487-w.
[18] F. José and A. V. Ingelmo, “What Do We Mean by GenAI? A Systematic Mapping of The Evolution, Trends, and Techniques Involved in Generative AI,” IJIMAI, vol. 8, no. 4, pp. 7–16, 2023.
[19] L. Petrillo, F. Martinelli, A. Santone, and F. Mercaldo, “Toward the Adoption of Explainable Pre-Trained Large Language Models for Classifying Human-Written and AI-Generated Sentences,” Electronics, vol. 13, no. 20, p. 4057, Oct. 2024, doi: https://doi.org/10.3390/electronics13204057.
[20] Y. Wang, Y. Pan, M. Yan, Z. Su, and T. H. Luan, “A survey on ChatGPT: AI-generated contents, challenges, and solutions,” IEEE Open J. Comput. Soc., vol. 4, pp. 280–302, 2023.
[21] Y. Wang et al., “GenAI Content Detection Task 1: English and Multilingual Machine-Generated Text Detection: AI vs. Human,” arXiv.org, 2025. https://arxiv.org/abs/2501.11012 (accessed Feb. 07, 2025).
[22] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," in Proceedings of the First Instructional Conference on Machine Learning, vol. 242, no. 1, pp. 29-48, Dec. 2003.
[23] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter," arXiv preprint arXiv:1910.01108, 2020.
[24] B. Sharimbaev and S. Kadyrov, “Automatic Language Identification from Audio Signals using LSTM-RNN,” 2023 17th International Conference on Electronics Computer and Computation (ICECCO), pp. 1–5, Jun. 2023, doi: https://doi.org/10.1109/icecco58239.2023.10146603.
[25] S. Kadyrov, C. Turan, A. Amirzhanov, and C. Ozdemir, “Speaker Recognition from Spectrogram Images,” 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST), pp. 1–4, Apr. 2021, doi: https://doi.org/10.1109/sist50301.2021.9465954.