Journal of Intelligent Systems and Internet of Things
Vol. 16, No. 01, PP. 152-165, 2025
DOI: https://doi.org/10.54216/JISIoT.160113

Human to Chatbot Text Classification Using Multi-Source AI Chatbots and Machine Learning Models

Mohammed Salah Ibrahim[1,*], Jabbar Abed Eleiwy[2], Hassan Mohamed Muhi-Aldeen[3], Yusra Al-Yasiri [4], Ahmed Adil Nafea[1]

[1]Department of Artificial Intelligence, College of Computer Science and IT, University of Anbar, Ramadi, 3100, Iraq

[2]Department of Applied Sciences, University of Technology-Iraq, 52 Alsena str., Baghdad, 10053, Iraq

[3]Department of Computer Engineering, Aliraqia University, 22 Sabaabkar, Adamia, Baghdad, 10053, Iraq

[4]Department of Kindergarten and Special Education, Aliraqia University, 22Sabaabkar, Adamia, Baghdad, 10053, Iraq

Emails: Moh.salah@uoanbar.edu.iq; jabar.a.eleiwy@uotechnology.edu.iq; muhialdeen.hassan@aliraqia.edu.iq; yusra.h.naser@aliraqia.edu.iq; ahmed.a.n@uoanbar.edu.iq

Abstract

The fast growth of artificial intelligence technologies, especially language processing technology has obscured the lines in between human-generated text comparing to chatbot-generated message. Recognizing which generated such, a text is essential for applications like information generating and manipulated text in order to guarantee authenticity between communicated parties. This research applies to a set of machine learning models to identify text as either human-written or chatbot-generated. The methodology of this research starts with a dataset including text generated from different Large Language Models (LLMs) along with a text generated by a human. After that, Tf-Idf ranking vectorization was used to define word embedding has and represent the text numerically. Then, different Machine Learning (ML) models leveraged recognize whether a human or a chatbot generated a text. The ML models applied include Logistic Regression, Random Forest, Decision Tree, Gradient Boosting, Naïve Bayes, and XGBoost. For this study accuracy, precision, recall, F1-score were used to evaluate the system. The dataset first was split into 80% for training and 20% for testing. Out of all implemented models, the Random Forest model reported the best with accuracy of 88%. Logistic Regression reported a close accuracy of 85%. The Random Forest model showed an 8% improvement compared to previous studies that reported an accuracy of 80%. Confusion matrices revealed that the Random Forest model provided high precision and recall, minimizing classification misleading of human or chatbot text. The research focused on studying the ability of ML models in identifying human vs. chatbot-generated text. The results showed the RF model was the best among other models with 88% accuracy. This accuracy shows a possible usage of such models in real-world applications that requires the confidentiality of human writing.

Received: October 19, 2024 Revised: January 11, 2025 Accepted: January 31, 2025

Keywords: Chatbot; Text Classification; Artificial Intelligence; Machine Learning

1. Introduction

The new development in AI technologies with the advancements of transformers and the text embedding techniques made it more difficult to compare human-generated to chatbot-generated text. Chatbots are natural language conversational bots developed using LLMs. Example of such chatbot is the ChatGPT built by OpenAI and the Google Gemini chatbot. These bots provide content generation services as well as conversational services. Despite the benefit these chatbots provide, they raise concerns about authenticity of content generation. It becomes hard to tell whether chatbot or a human generates a text. It is required for applications like AI misinformation generation, fake news and many others [1].

The LLMs receive a prompt from the user as input. Then, LLMs predict the next word in the context of the prompt according to content that its model trained on. LLMs trained on very large datasets of articles, books, news, and many more text. Due to that, such models can generate creative outputs such as whole answers and sentences or even poems and essays. These models developed depending the concept of transfers and attention mechanism [2], [3]. Model like GPT-4 has been trained on a large amount of text. Due to that such model was able to pass many professional and academic tests [4].

Users interact with an LLM via AI conversational chatbots. An AI chatbot is an application leverages LLMs technology to provide human-like conversation [5]. The old versions of chatbots used rule-based mechanism which response according to a set of answers they have set to. Then, these chatbots developed with AI-powered advancement using ML and Natural Language processing NLP too that helped understanding and interpreting user input according to its contexts, sentiments, and intentions [6], [7]. Figure 1. shows the general architecture of AI conversational chatbot [8].

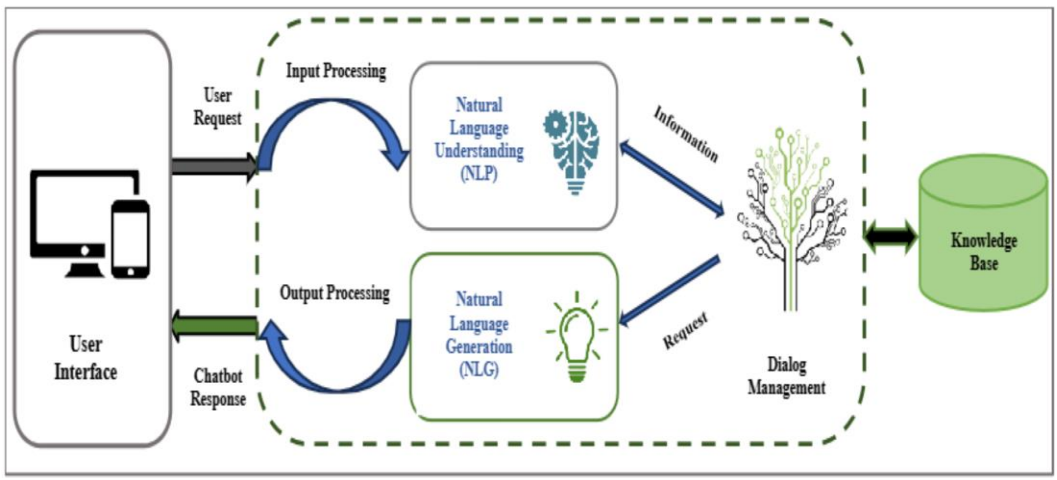


Figure 1. AI conversational chatbot architecture [8]

The OpenAI technical report regarding their latest LLMs model called GPT-4 reported that despite GPT4's advancement in content generating, it still producing errors in reasoning and hallucinating. Their report recommended full care to be taken regarding output chatbot provide. Although, GPT-4 considerably decreased delusions compared to previous models, it still not accurate enough to be considered as credible source of information. According to its evaluation on different factuality evaluation dataset, its accuracy was around 80%, which is better than previous models 19% scores higher than others [4]. Figure 2 shows GPT4 performance on different factual datasets.

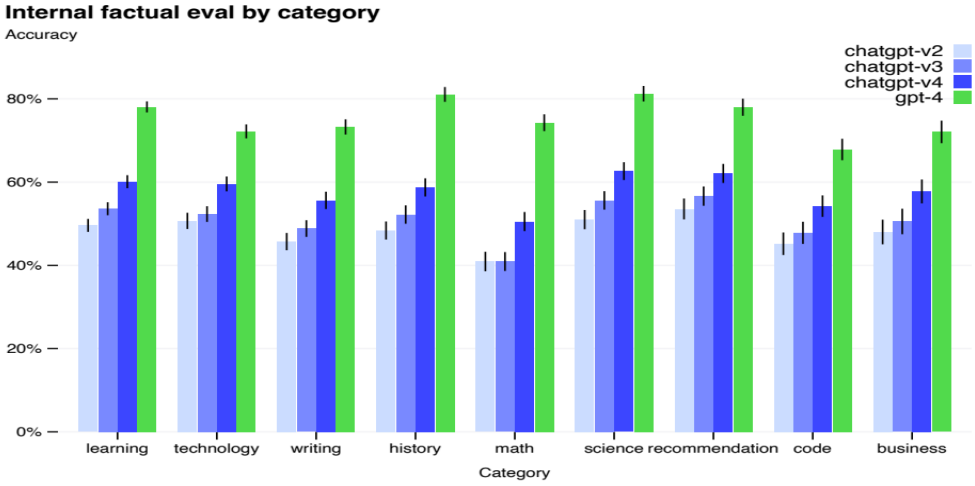


Figure 2. GPT-4 performance comparing to previous versions of GPT on nine adversarial-factuality datasets [4].

In a study presented by Zhuo et al. about the ethical issues related to ChatGPT and many advanced large language models LLMs, they reported that despite LLMs have influenced numerous technologies, they still produce have biasness and toxicity that pose moral and societal risks. Through many experiments, authors measured ethical issues such as dependability, biases, dependability, robustness and toxicity using many datasets. Their study ended with the fact that current models still have many ethical concerns. The authors emphasized the importance of their findings on ChatGPT's AI ethics and offered potential future concerns and practical design considerations for LLMs. Ultimately, the authors concluded that their research provides valuable insights into future endeavors to identify and mitigate ethical risks associated with LLM applications[9]. A lot of other studies presented to measure accuracy, biasness, toxicity, authenticity, and trustworthiness of LLMs chatbots such as [10], [11], [12], [13], [14]. These studies drive our attention to the importance of determining whether a text generated by a human or AI chatbot to guarantee authenticity and credibility.

Machine learning models have emerged as a solution for text classification. They showed great help because they not only reduce the time for processing but also their ability to analyze huge volumes of text data in a fast way. Standard models like Logistic Regression and Decision Trees have been successfully applied to text classification problems. Furthermore, models like Random Forest and gradient boosting techniques, such as XGBoost, have gained popularity for their ability to capture non-linear relationships in complex datasets.

The process of distinguishing human vs. chatbot text requires extracting meaningful and linguistic patterns and from the text. Human-written text often has the property of a higher degree of creativity, irregularity, and context-dependent distinctions. On the other hand, chatbot-generated text leans toward being more structured and repetitive with high formal writing and styling. This is because chatbots reflect the inherent biases of the training data and algorithms used to generate it [15]. Using proper feature detection methods like Term Frequency-Inverse Document Frequency (Tf-Idf) in addition to applying selected machine learning technique, AI chatbot writings can be detected with high accuracy.

This research investigates the effectiveness of machine learning models in identifying a human from AI chatbot text. Different ML models were applied in this research including Logistic Regression, Random Forest, Decision Tree, Gradient Boosting, Naïve Bayes, and XGBoost. The project aims to find an effective model for this task. The main goal is to help with future studies as well as applications involved in detecting human and AI writing apart.

2. Related Work

Text classification has been studied extensively, but with the emergent of AI-powered chatbot, this field returned as a trend in research area.

Godghase et al [16] build a machine learning model to distinguish human written text from chatbot generated text. The authors developed a large dataset consisting of more than 750,000 human written text along with chatbot-generated text for the same topic. In their research, authors used feature analysis along with embedding is methods to train the ML model. In the feature extraction step, they focused on a group of linguistic characteristics represented by word choice and sentence structure. In addition to that, they used PCA and LDA for feature engineering. Traditional embedding methods like GloVe, Word2Vec, and Tf-Idf and transformer-based technique like BERT were applied used to build text representation numerically. With different feature extraction and selection methods and different ML models, they achieved high classification accuracy of 96%.

Katib et al. [17] developed a novel method to detect texts from humans and ChatGPT through the Tunicate Swarm Algorithm among with Long Short-Term Memory Recurrent Neural Network (TSA-LSTMRNN). The approach used Tf-Idf, word embedding's along with count vectorizers to extract features. The LSTMRNN model handled the process of training and classification of text. A TSA helps pick the best settings to improve detection quality. Tests on reference data sets proved the TSA-LSTMRNN system outperformed other recent methods with high accuracy reached to 93.17 % for human texts and 93.83 % for ChatGPT text.

Luo et al. [18] conducted an experiment with over 62 thousand customers to examine the performance of chatbots and human in outbound income calls. The results of their study showed that chatbots are effective as human expert in that field and they are better than laymen four times in providing income calls and increasing purchases. However, revealing the chatbot identity to people lowered the purchase rate to 79.7%. The disclosure also reduced name duration, as clients perceived disclosed chatbots as much less informed and empathetic. This terrible effect of disclosure may be mitigated with the aid of delaying the chatbot's identity monitor and leveraging clients' previous AI revel in. The look at offers precious insights for chatbot software, client targeting, and conversational commerce strategies.

Mitrović et al. [19] investigated the difficulty of distinguishing human writings from ChatGPT writings, especially with short text reviews. The study used an explainable AI tool called Shapley Additive explanations (SHAP) to explain the ability of ML in distinguishing chatbot from human text. A fine-tuned transformer-based model used to classify the text. The study found it is hard to discover alternated or rephrased ChatGPT text. Their model was able to distinguish human from ChatGPT text with 79% accuracy. The investigations revealed that ChatGPT as a chatbot provides text that is more impersonal and polite with emotions in expressions with using more complicated vocabulary that is different when compared to human writing.

Akpan et al. [20] examined the evolving role of Conversational and Generative AI (CGAI/GenAI), particularly in human-chatbot interactions, and its impact on various fields. Their study highlights the rapid growth of CGAI research, with 96% of publications occurring between 2019 and 2023. Key use cases of CGAI, such as ChatGPT, are prominent in education, particularly in computer science, healthcare, engineering, and business. The study identifies significant benefits, like enhanced human-computer interaction and support in academic tasks (e.g., syllabus creation, testing, and writing). However, concerns about misuse, such as plagiarism and privacy violations, especially in healthcare, were also noted. The authors emphasize the need for strategies, policies, and detection mechanisms to address potential challenges in both educational and operational contexts, advocating for further research into CGAI integration in decision-making systems.

Prova et al. [21] focused on the growing challenge of distinguishing between AI-generated and human-written text, an issue with significant ethical, legal, and social implications. Their study proposed an AI detection model using ML techniques like XGB Classifier, SVM, and the BERT

deep learning architecture. The results showed that BERT outperformed the other models, achieving an accuracy of 93%, compared to 84% for XGB and 81% for SVM.

Repaka et al. [22] addressed the demanding situations of computational efficiency and aid allocation in training LLMs. Their examine utilizes the "Human vs. LLM Text Corpus" from Kaggle to evaluate allotted training strategies along with data parallelism, model parallelism, and pipeline parallelism. The assignment targets to both distinguish among human- and AI-generated texts and optimize the schooling technique for LLMs. Preliminary outcomes display an 80% accuracy in classifying texts and a 2.1x increase in education time performance the usage of parallelism techniques. The observe contributes to the expertise of scalable LLM education, with implications for improving efficiency, accuracy, and fault tolerance in disbursed gadget getting to know structures.

Our research using the same dataset used by the last research [22]. This research uses this comprehensive dataset, which incorporates text generated by popular chatbots like GPT-3.5, GPT-4, Claude and Gemini-Pro and other AI models such as Bloom-7B, Falcon-180B, Flan-T5, Goliath-120B, and LLaMA-13B. This variety of chatbot sources enables deeper analysis of variances between texts generated by human AI-generated one. Previous studies examined text from ChatGPT or single platforms. Our research uses text across different platforms versus text written by humans, which makes our work distinguished from other previous studies.

3. Methodology

This study employs a machine learning-based approach to classify text as either human-written or chatbot-generated. The methodology consists of four main stages: dataset preparation, feature extraction, model training and evaluation, and performance comparison. Figure 3. Shows the main stages of our proposed system.

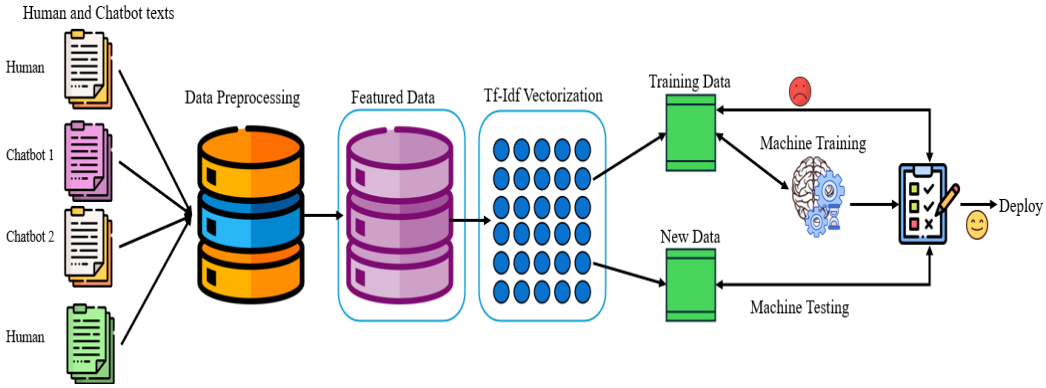


Figure 3. Methodology of our proposed model.

3.1 Dataset

The dataset used for this study is the publicly available "HumanvsChatbot" dataset from Kaggle[23]. This dataset has more than 7500,000 entries. It has five attributes represented by such as text (human or AI chatbot), source (written by human or AI chatbot), prompt ID, text length, and word count. Figure 4 shows the distribution of text generated by different sources. Approximately 44% of the texts originate from human sources; while 63 different AI models like GPT-3.5, GPT-4, Claude and Gemini-Pro and other AI models such as Bloom-7B, Falcon-180B, Flan-T5, Goliath-120B, and LLaMA-13B, generate the remaining 66%. This variety of chatbot sources enables deeper analysis of variances between texts generated by human AI-generated one. In this study, human-generated text was labeled as 1, while chatbot-generated text from all other sources was labeled as 2. This binary classification task aims to distinguish between these two categories based on linguistic features. In this dataset, there are around 56% of the text generated by humans and 44% generated by AI chatbots. Figure 4. Shows that distribution clearly.

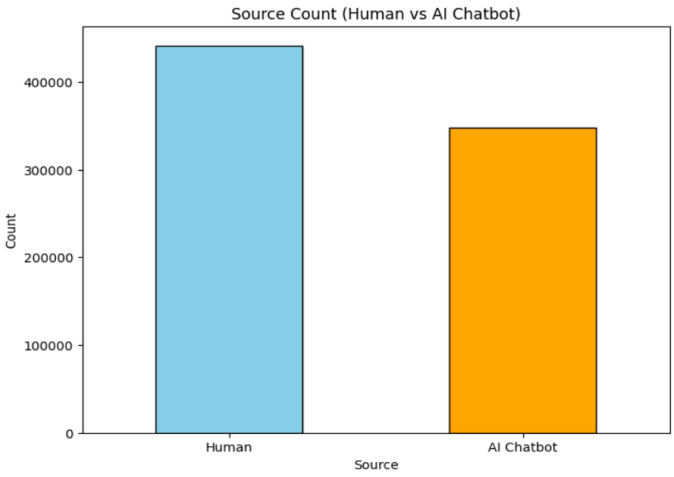


Figure 4. Text source distribution as human or AI chatbot

3.2 Data Preprocessing

AI-powered chatbots are restricted to what they are trained on and learned from. They have processed formal and informal styles of writing. They learned not only from internet pages, but also many other sources such as books, articles, news and many others. We tried to keep text generated by both human and chatbots as it is so that pattern can be discovered clearly. We did not remove stop words. We also did not apply stemming or lemmatization. Stemming or lemmatization can do over-normalization like turning distinct words into their basic form that give another meaning like "organization" and "organize", which lower the discriminative power of features extraction using Tf-Idf.

Moreover, we did not apply any filtering or removing techniques such as word length and word count. All that to keep human and chatbot styling in its real form to increase pattern identification. However, we removed special characters, extra whitespace, and non-alphanumeric tokens, and lowercased all text for consistency. In addition, we added bigrams as a choice to the feature extraction step. Bigrams help with capturing word pairs that can provide contextual relationships between words and help identifying semantic importance.

3.3 Feature Extraction

In order to define every word, represent it as a vector so that the machine can understand it, and study its patterns, *Tf-Idf* (Term Frequency-Inverse Document Frequency) vectorization was applied to convert the text data into numerical representations. In order to get a good representation for each gram or bigram, the maximum feature dimension was defined at 5000 to represent the most relevant terms to current term. To balance computational efficiency and model performance. The *Tf* represents the occurrence of a word *w* in a text *t* compared to the number of words in that text (See Equation 1). The *Idf* measures the rareness of the *w* according to the whole dataset texts (See Equation 2). The score of the *Tf-Idf* represents the weight of the *w* in its *t* (See equation 3). The higher the score, the more important the *w* to its *t*.

$$Tf = \frac{\text{occurrence of a } w \text{ in a } t}{total\ t\ words} \quad (1)$$

$$\text{Idf} = \log\left(\frac{total\ of\ ts\ in\ a\ dataset}{number\ of\ ts\ contain\ w}\right) \quad (2)$$

$$Tf - Idf = Tf \times Idf \quad (3)$$

Where t is the text in the dataset and w is a word in a text t.

3.4 Data Splitting

After preparing the dataset and getting all the required features, the dataset was split into training and testing. For this work, 20% of the dataset kept for testing and the remaining dataset was used to train the machine-learning model. The target label, which is the source of the text, human (labeled as 1) or chatbot (labeled as 2), were saved for measuring model performance in predicting text whether generated by a human or a AI-powered chatbot.

3.5 Machine Learning Models

In this study, six ML models implemented include LR, RF, NB, GB, DT and XGBoost. These models are commonly used for classification tasks:

3.5.1 Random Forest

RF is an ensemble learning technique that aggregates the predictions of multiple DT. Each tree is trained on a random subset of the data, and a random subset of features is considered at each split. The model predicts the class by taking a majority vote from the trees.

The Gini impurity criterion used to split the nodes in each tree is given by:

$$\text{Gini}\ (t) = 1 - \sum_{i=1}^{c} p_i^2 \qquad (4)$$

where c is the number of classes, p_i is the probability of class i in node t.

3.5.2 Logistic Regression

LR is a linear model used for binary classification. It predicts the probability that a given input belongs to a particular class, based on the logistic function. The equation for LR is:

$$p(\mathcal{Y} = 1 + | X) = \frac{1}{1+e^{-(X\beta)}} \qquad (5)$$

Where X is the feature vector of the input text (the TF-IDF representation), β are the model coefficients, P(y=1|X) is the probability that the text is chatbot-generated.

3.5.3 Gradient Boosting

GB is an ensemble learning technique that builds trees sequentially, with each tree trying to correct the errors made by the previous one. The objective function for GB is the residual sum of squares (RSS) or other loss functions depending on the task. For classification tasks, the gradient of the loss function is minimized iteratively.

The model predicts the output $\hat{y}t$ for each instance at iteration $t\ as$:

$$\hat{y}t = \hat{y}t - 1 + \eta . h_t(X) \qquad (6)$$

$where\ \hat{y}t - 1$ is the prediction from the previous iteration, (X) is the newly added tree's prediction at iteration t, η is the learning rate. Each subsequent tree $h_t(X)$ fits the residuals of the previous model, minimizing the error iteratively.

3.5.4 Decision Tree

DT recursively partition the feature space to classify data points. At each node, the feature that maximizes information gain is selected. Information gain is calculated using the entropy measure, where the entropy $H(S)$ *for a set* S of data.

$$H(S) = -\sum_{i=1}^{C} p^i \, log_2(p_i) \qquad (7)$$

$where\ C$ is the number of classes, p_i is the proportion of class i in the dataset S.

A split is made on the feature that provides the greatest reduction in entropy. The DT recursively repeats this process, splitting the data at each node, until it reaches a predefined maximum depth or all data in a leaf node are of the same class.

3.5.5 Naive Bayes

NB is a probabilistic classifier based on Bayes' Theorem, assuming conditional independence of features. The posterior probability of class y given features x $(x_1, x_2, x_3, \ldots . x_n)$

$$p(y \mid x) = \frac{p_{(y)} \prod_{i=1}^{n} p(x_i \mid y)}{p(x)} \quad (8)$$

Where P(y) is the prior probability of class, $p(x_i \mid y)$ is the likelihood of feature x_i given class y, P(X) is the evidence, which normalizes the probabilities.

The class with the highest posterior probability is selected as the predicted class.

3.5.6 XGBoost

It is an optimized implementation of GB that uses DT as weak learners. XGBoost is known for its speed and performance, leveraging regularization and efficient computation. The objective function in XGBoost is a combination of a loss function L and a regularization term Ω.

$$\iota(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{k} \Omega(f_k) \quad (9)$$

$where\ L(y_i, \hat{y}_i)$ is the loss function for the $_i$ th instance, $\Omega(f_k)$ is the regularization term for tree f_k to control model complexity.

The model minimizes this objective function iteratively by adding trees that reduce the residuals of previous predictions.

3.6 Model Evaluation

The models are evaluated based on the following performance metrics:

1. Accuracy: The proportion of correctly classified instances:

$$accuracy = \frac{TP + TN}{Total\ Instances} \quad (10)$$

2. Precision: The proportion of TP among all cases predicted as positive:

Precision = $\frac{TP}{TP+FP}$ (11)

3. Recall: The proportion of TP among all actual positives:

Recall = $\frac{TP}{TP+FN}$ (12)

4. F1-score: The harmonic mean of precision and recall:

F1-score = $2*\frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$ (13)

4. Results and Discussion

For this research implementation, Kaggle platform was used with its standard resources explained in Table 1. The models were programed and run using Python 3.10.12 environment. Many libraries were used like nltk, scikit-learn and many others.

Table 1: Kaggle computational resources setting.

Resource	Setting
GPU	NVIDIA Tesla P100
GPU Ram	16 GiB
Disk Space	57.6 GiB

Table 2. Reports the parameter setting for the whole models applied. The standard setting was reported. We did not apply any parameter tuning technique and left that for the future work.

Table 2: Parameter setting for every machine-learning model applied.

Model	Parameter Setting
Random Forest Classifier	Number of decision trees=100, maximum depth=None
Logistic Regression	Objective function solver= Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm, maximum iterations=1000
Gradient Boosting	Learning rate=0.1, Number of decision trees=100, maximum depth=None
Decision Tree	Criterion='gini', maximum depth=None
Naive Bayes (MultinomialNB)	Alpha=1.0, fit prior=True
XGBoost	Number of boosting rounds estimated=100, learning rate=0.1, maximum depth=6, subsample=0.8, colsample_bytree=0.8, evaluation metric='mlogloss'

The experiment implemented over the seven machine learning models including Logistic Regression (LR), Gradient Boosting (GB), Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), and XGBoost (XGB). Table 3. Shows the performance of these six models. Out of these seven models, the Random Forest model reported the highest accuracy of 88% and the same for F1-score, which made it the best model in identifying whether a human wrote a text or chatbot text. Logistic Regression model was the next best model among the others. Naïve Bayes and Decision Tree models reported the worse results compared to all other models. The reason that these models did not perform well might be to the degree of correlation between words that model could not discriminate between human words and chatbot words, which led to bad performance.

Table 3: Models' performance using precision, recall, F1-score, and accuracy

Model	Precision (Weighted)	Recall (Weighted)	F1-Score (Weighted)	Accuracy
Logistic Regression	85	85	85	85
Gradient Boosting	83	82	81	82
Decision Tree	80	80	80	80
Random Forest	**88**	**88**	**88**	**88**
Naive Bayes	76	76	76	76
XGBoost	85	84	84	84

In order to discuss the performance of these models according to class level classification, a heatmap was developed (See Figure 5). The heatmap shows an important variation between models in terms of label classification. Class 1 represents human text and class 2 represents AI chatbot text. In general, Random Forest reported the highest for class 1 with precision of 92%, recall 80%, and F1-score of 88%. It also reported the highest for class 2 with 86% precision, 0.94 recall, and 0.92 for F1-score. Logistic Regression also did well across the two class. On the other hand, Naive Bayes reported the worst over all models, particularly for Class 1, where it achieves the lowest precision 75%, recall 69%, and F1-score of 72%.

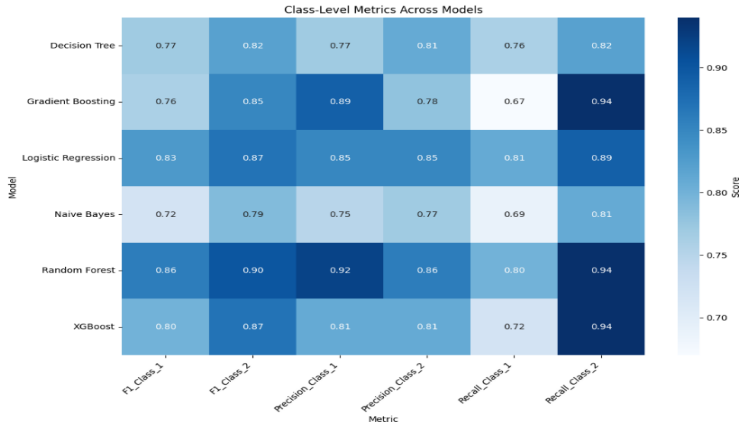
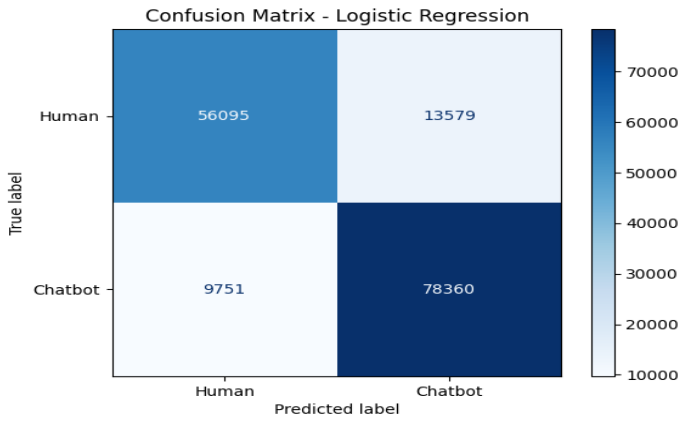


Figure 5. A heatmap of class level models' evaluation

Despite the fact that 56% of the text in the dataset was human generated text and 44% of the text was AI chatbot generated text, this imbalance did not effect on class 2, AI Chatbot, which reported higher precision, recall, and F1-scores across most models. These results suggest that the models were able to recognize many patterns or linguistic features with chatbot-generated text compared to human text. This also leads to the fact that AI chatbot, according to what they have learned and trained on, have consistent stylistic or structural characteristics that make them different from human in writing. On the other hand, human text has diverse styling, and it is more nuanced, which can lead to complexity for models to grasp patterns to recognize human writing (Class 1).

The performance can be seen clearly using confusion matrices. For each model, a confusion matrix was generated to analyze their ability to minimize misclassification. Figure 6 shows the confusion matrix for some of the implemented models.



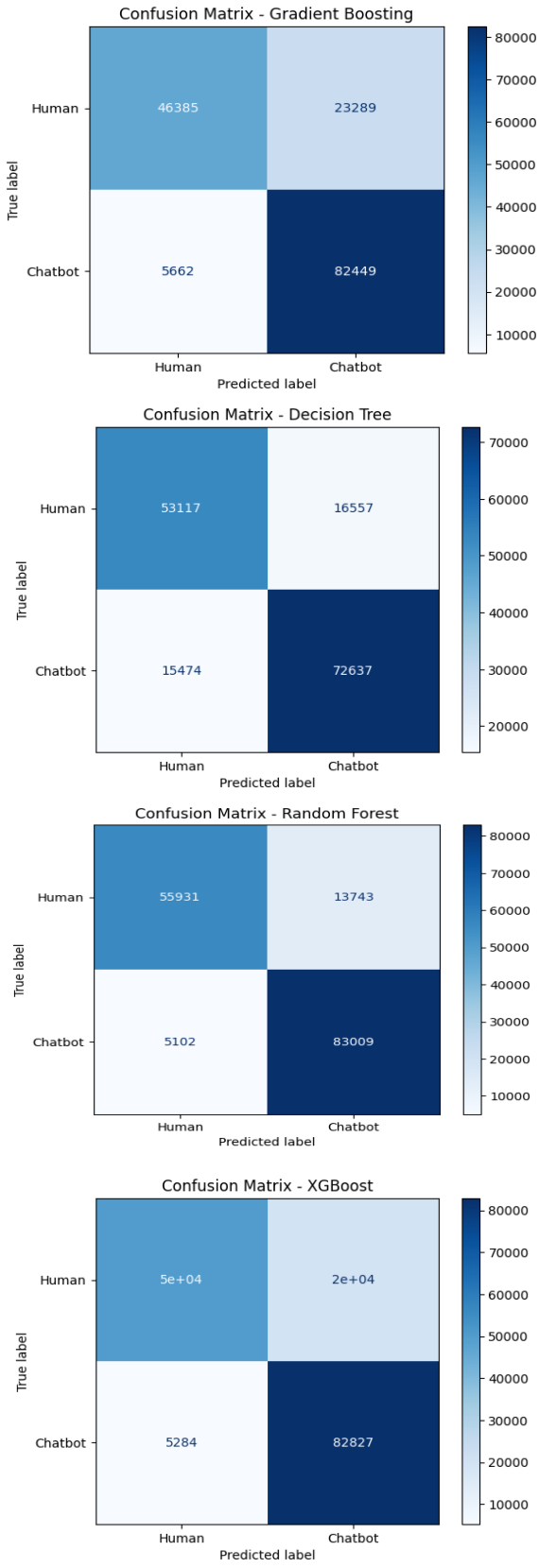


Figure 6. Confusion matrix for all of the implemented models.

From the results above, it can be concluded that selecting the model for such a task is an important step. The Random Forest model shows the highest performance compared to other models. It shows this model ability to handle high high-dimensional data. Random Forest as an ensemble model was able to average predictions from multiple decision trees making it particularly suitable for such a task. Despite such a model was used in other studies for the same dataset such as [22], however, they did not report that much accuracy. The highest accuracy reported from this previous study was 80% while in this research, it is 88% and that is 8% improvement. Our results showed performance better than they were in the way we treated data. In their case, they delt with 33 class, but in our case, we considered all AI chatbot text as one class and human text as another class. This provided a high amount of data that helped the machine-learning model to recognize text and classify it. On the other hand, a study conducted by Mitrović et. al. [19] reported low accuracy (79%) on another datasets. In another research that used the GPTZero to distinguish human from chatbot writing which reported low accuracy at 80% [24].

5. Conclusion

This study focused on classifying human-generated and chatbot-generated text using machine learning. The study used the "HumanvsChatbot" dataset from Kaggle. The text was first preprocessed and then features were extracted using the TF-IDF vectorization. Different machine learning models were implemented. After training and testing, the Random Forest model showed the highest performance among the other models with an accuracy of 88% and an F1-score of 0.88. Such a model was able to recognize a text belonging to a human or AI chatbot returns to its ability in dealing with high dimensional data and extracting patterns from it. This study showed the importance of model selection and feature engineering in achieving high accuracy in text classification. For future work, we recommend applying deep learning models that use transformers mechanisms such as BERT or GPT. Another suggestion is to apply any of the optimization algorithms to find the best model setting.

References

[1] E. Lozić and B. Štular, “Fluent but Not Factual: A Comparative Analysis of ChatGPT and Other AI Chatbots’ Proficiency and Originality in Scientific Writing for Humanities,” *Future Internet*, vol. 15, no. 10, Art. no. 10, Oct. 2023, doi: 10.3390/fi15100336.

[2] K. I. Roumeliotis and N. D. Tselikas, “ChatGPT and Open-AI Models: A Preliminary Review,” *Future Internet*, vol. 15, no. 6, Art. no. 6, Jun. 2023, doi: 10.3390/fi15060192.

[3] A. Vaswani et al., “Attention Is All You Need,” Jul. 23, 2023, arXiv: 1706.03762. [Online]. Available: https://arxiv.org/abs/1706.03762. doi: 10.48550/arXiv.1706.03762.

[4] OpenAI et al., “GPT-4 Technical Report,” Mar. 04, 2024, arXiv: 2303.08774. [Online]. Available: https://arxiv.org/abs/2303.08774. doi: 10.48550/arXiv.2303.08774.

[5] E. Adamopoulou and L. Moussiades, “An Overview of Chatbot Technology,” in *Advances in Intelligent Systems and Computing*, vol. 1056, Cham, Switzerland: Springer, 2020, pp. 383–396, doi: 10.1007/978-3-030-49186-4_31.

[6] B. Galitsky, “Adjusting Chatbot Conversation to User Personality and Mood,” in *Artificial Intelligence for Customer Relationship Management*, Human–Computer Interaction Series, Cham, Switzerland: Springer, 2021, pp. 93–127, doi: 10.1007/978-3-030-61641-0_3.

[7] Z. Peng and X. Ma, “A survey on construction and enhancement methods in service chatbots design,” *Springer*, vol. 1, no. 3, pp. 204–223, 2019. [Online]. Available: https://link.springer.com/article/10.1007/s42486-019-00012-3.

[8] S. Izadi and M. Forouzanfar, “Error Correction and Adaptation in Conversational AI: A Review of Techniques and Applications in Chatbots,” *AI*, vol. 5, pp. 803–841, Jun. 2024, doi: 10.3390/ai5020041.

[9] T. Y. Zhuo, Y. Huang, C. Chen, and Z. Xing, “Red teaming ChatGPT via Jailbreaking: Bias, Robustness, Reliability and Toxicity,” May 29, 2023, arXiv: 2301.12867. [Online]. Available: https://arxiv.org/abs/2301.12867. doi: 10.48550/arXiv.2301.12867.

[10] Y. Huang et al., “TrustLLM: Trustworthiness in Large Language Models,” Sep. 30, 2024, arXiv: 2401.05561. [Online]. Available: https://arxiv.org/abs/2401.05561. doi: 10.48550/arXiv.2401.05561.

[11] T. R. Hannigan, I. P. McCarthy, and A. Spicer, “Beware of botshit: How to manage the epistemic risks of generative chatbots,” *Elsevier*, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0007681324000272.

[12] M. U. Hadi et al., “Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects,” Nov. 16, 2023, [Online]. Available: https://techrxiv.23589741.v4. doi: 10.36227/techrxiv.23589741.v4.

[13] S. Wyer and S. Black, “Algorithmic bias: sexualized violence against women in GPT-3 models,” *AI Ethics*, Jan. 2025, doi: 10.1007/s43681-024-00641-0.

[14] T. Choudhary, “Political Bias in Large Language Models: A Comparative Analysis of ChatGPT-4, Perplexity, Google Gemini, and Claude,” *IEEE*, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10817610/.

[15] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event Canada, ACM, Mar. 2021, pp. 610–623, doi: 10.1145/3442188.3445922.

[16] G. A. Godghase, R. Agrawal, T. Obili, and M. Stamp, “Distinguishing Chatbot from Human,” arXiv: 2408.04647v1, Jan. 19, 2025. [Online]. Available: https://arxiv.org/abs/2408.04647v1.

[17] I. Katib, F. Y. Assiri, H. A. Abdushkour, D. Hamed, and M. Ragab, “Differentiating Chat Generative Pretrained Transformer from Humans: Detecting ChatGPT-Generated Text and Human Text Using Machine Learning,” *Mathematics*, vol. 11, no. 15, Art. no. 15, Jan. 2023, doi: 10.3390/math11153400.

[18] X. Luo, S. Tong, Z. Fang, and Z. Qu, “Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases,” *Marketing Science*, vol. 38, no. 6, pp. 937–947, Nov. 2019, doi: 10.1287/mksc.2019.1192.

[19] S. Mitrović, D. Andreoletti, and O. Ayoub, “ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-generated Text,” arXiv.org. [Online]. Available: https://arxiv.org/abs/2301.13852v1.

[20] I. J. Akpan, Y. M. Kobara, J. Owolabi, A. A. Akpan, and O. F. Offodile, “Conversational and Generative Artificial Intelligence and Human–Chatbot Interaction in Education and Research,” *Int. Trans. Oper. Res.*, vol. 32, no. 3, pp. 1251–1281, 2025, doi: 10.1111/itor.13522.

[21] N. Prova, “Detecting AI Generated Text Based on NLP and Machine Learning Approaches,” Apr. 15, 2024, arXiv: 2404.10032. [Online]. Available: https://arxiv.org/abs/2404.10032. doi: 10.48550/arXiv.2404.10032.

[22] K. T. Repaka, M. A. Bondugula, and S. S. Adibhatla, “Benchmarking Distributed Machine Learning Systems with Large Language Models on Human vs. LLM Text Corpus,” Jan. 19, 2025. [Online]. Available: https://disml2024.github.io/disml-workshop-2024/assets/8_945276_86359389_Group8_DISML_Project_Report.pdf.

[23] Z. Grinberg, “Human vs. LLM Text Corpus,” *Kaggle*, Jan. 30, 2025. [Online]. Available: https://www.kaggle.com/dsv/7378735.

[24] F. Habibzadeh, “GPTZero Performance in Identifying Artificial Intelligence-Generated Medical Texts: A Preliminary Study,” *J. Korean Med. Sci.*, vol. 38, no. 38, p. e319, Sep. 2023, doi: 10.3346/jkms.2023.38.e319.