Comparative Analysis of Human and AI generated Text

Conference Paper · March 2024

DOI: 10.1109/SPIN60856.2024.10511301

Arti M.K.
Netaji Subhas University of Technology

2024 11th International Conference on Signal Processing and Integrated Networks (SPIN) | 979-8-3503-0843-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/SPIN60856.2024.10511301

Comparative Analysis of Human and AI generated Text

Shashank Kumar
Dept. of Electronics And Communication
Netaji Subhas University of Technology
New Delhi, India
shashankkr843@gmail.com

Sneha Tiwari
Dept. of Electronics And Communication
Netaji Subhas University of Technology
New Delhi, India
tsneha981@gmail.com

Rishabh Prasad
Dept. of Electronics And Communication
Netaji Subhas University of Technology
New Delhi, India
rishabh.prasad2003@gmail.com

Abhay Rana
Dept. of Electronics And Communication
Netaji Subhas University of Technology
New Delhi, India
abhayghps@gmail.com

Dr. Arti M.K.
Dept. of Electronics And Communication
Netaji Subhas University of Technology
New Delhi, India
arti_mk@yahoo.com

***Abstract*—AI (Artificial Intelligence) has emerged as a transformative tool that has revolutionized the things that we do daily. However, this rapid advancement in AI technology also raises ethical concerns that need to be addressed. This includes biases in the response of generative AI, interventions affecting the privacy of individuals, etc. Also, with easy access to emerging technologies like ChatGPT, Bard, DALL-E, etc., there has been a significant rise in the generation of fake content like fake images and deep-fake videos. Proper measures for the identification and validation of AI-generated content need to be established to minimize the circulation of false or fabricated information. Government regulations and public awareness are needed to ensure strict ethical practices in the content generated by AI. In this paper, a survey is conducted to collect human responses to a set of questions. The same questions are fed to AI tools to generate responses. These responses are then analyzed by various machine learning algorithms and studied on several parameters, like vocabulary richness, spelling errors, etc., to help detect whether the content is generated by AI or humans.**

***Index Terms*—ARTIFICIAL INTELLIGENCE, REAL VS FAKE, GENERATIVE AI**

I. Introduction

AI (Artificial Intelligence) is the branch of computer science that deals with the intelligence of machines that can perform tasks similar to human intelligence. [1] The central principles of AI revolve around reasoning, knowledge, planning, learning, communication, perception, and the ability to move and manipulate objects. One of the most important visionaries and theoreticians in the history of AI was Alan Turing (1912-1954), a British mathematician who in 1936 proved that a universal calculator (now known as Turing machine) is possible [2]. It suggests that a machine is capable of solving any problem as long as it can be represented and solved by an algorithm. AI is implemented in various ways as –

- Machine Learning: It is one of the applications of AI where the machine learns and improves from its experience automatically rather than explicitly programmed to perform certain tasks. Deep learning is a subset of machine learning which is based on ANN (Artificial Neural Networks). [3] Based on the methods and way of learning, there are various types of machine learning algorithms such as unsupervised learning, supervised learning, semi-supervised learning, and reinforcement learning.
  In unsupervised learning, the machine is trained on an unlabelled dataset, while on the other hand, a labelled dataset is considered in supervised learning. Semi-supervised learning algorithm lies between these two, i.e., the dataset consists of a few labels but mostly unlabelled data. Reinforcement learning is based on a feedback mechanism. In this type of learning, an AI agent automatically explores its surroundings by the hit and trial method, takes actions, learns from the experiences, and improves the overall performance.
- NLP (Natural Language Processing): [4] It is the field of AI that aims to develop algorithms that enable computers to understand, interpret, generate, and manipulate human languages.
- Neural Networks: It is an artificial system inspired by biological neural networks. It consists of a huge connected network of computational neurons organized in layers. By adjusting the weights of the network, neural networks can be trained to approximate virtually any non-linear function to a certain degree of accuracy.

In today's modern world, AI finds it's application in almost every sector from using it in healthcare to make better and faster diagnosis to using it in agriculture for crop monitoring and predictive analysis. But, since every coin has two sides, so is the case with the revolution of AI. [5] Since, AI is helpful in automating the tasks, low-skilled and routine jobs are the most susceptible to be displaced by the AI. McKinsey Global Institute (2018) studied the future of work and found that though this revolution of AI will create new opportunities for employment, it majorly depends on the speed of technological advancement and the ability of workers to adapt to new technologies.

[6] Generative AI, as a concept, encompasses computational methods that can create fresh and meaningful content, including text, images, or audio, derived from training data. The adoption of this technology, facilitated by tools like GPT-4, DALL-E, Bard AI, GitHub Copilot, and others, has brought about a transformative shift in our work processes.

979-8-3503-0843-3/24/$31.00 ©2024 IEEE

But, the ease of accessibility to these tools has also raised ethical concerns. AI is now so much powerful that it has become difficult for humans to detect whether the content is generated by AI or human. An illustrative experiment in the form of a game of trivia was conducted to analyse how efficiently can humans detect whether the content generated is by the AI or human [7]. It has shown that there is 50% chance of correct detection of the content which is an alarming situation as AI can be misused to generate false [8] and provoking content [9] to spread hate in the community. [10] The rise of deepfake images and videos specially in the area of adult movies and politics has imposed serious threats. [11] The ethical considerations of DALL-E involve concerns about bias and discrimination, privacy, and unintended outcomes. For instance, the images it creates could uphold unrealistic beauty standards, strengthen gender stereotypes, or add to the objectification of women.

To address the issues posed by the generative AI tools, this paper aims to analyse the responses generated by AI when compared with responses of humans using various machine learning algorithms on several parameters like vocabulary richness, spelling errors, grammatical score, readability score, etc.

II. Proposed Methodology

Looking at the past work, various activities have been conducted for the participants to differentiate between human and AI-generated content. The Turing test [12] is one of the most popular tests for this classification. In this test, an interrogator asks questions from both machines and humans, and based on the response they have to identify whether the given response is AI or human-generated. However, the methodology here aims to identify how the machine classifies the two contents and what are their identification scores.

A. Data Collection

For data collection, a survey of 10 questions based on Google Forms is conducted. The questions are formulated such that people can either have varied personal experiences as a response or general responses. Some of the questions asked in the survey are as follows:

- Is the trend of people watching only VFX-heavy or high-budget movies irrespective of the storyline in theatres good for the film industry? Explain why/why not?
- Do you think conducting space missions like Chandrayaan-3, Aditya L-1, etc. is a waste of money? Why/Why not? Explain.
- What according to you can be a person's ideal weekend plan?

A total of 700 text responses were collected. However, the data collected is from a specific demographic location in the English language (which is not the primary language for most of those who responded). Following the collection of human responses, about 600 responses are generated using various AI tools like ChatGPT, Bard, and Aichatting for the dataset.

B. Content Selection and Curation

An equal number of human and AI-generated responses were selected to create a balanced dataset. To minimize the bias between human and AI-generated content data is curated keeping the following factors in mind:

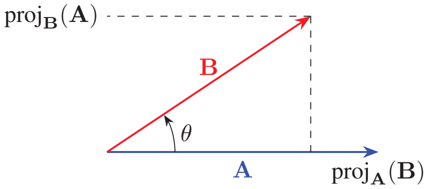
$\mathrm{proj}_{\mathbf{B}}(\mathbf{A})$
$\mathbf{B}$
θ
$\mathbf{A}$
$\mathrm{proj}_{\mathbf{A}}(\mathbf{B})$

Fig. 1. Illustration of Cosine Similarity. This figure demonstrates the vectors $\mathbf{A}$ and $\mathbf{B}$, the angle θ between them, and the projections $\mathrm{proj}_{\mathbf{A}}(\mathbf{B})$ and $\mathrm{proj}_{\mathbf{B}}(\mathbf{A})$. The cosine of angle θ gives a cosine similarity score between two vectors.

- Responses that are of less than 50 words were removed from the dataset.
- Human responses that were casual or highly unrelated were removed from the dataset.
- To generate AI responses, different prompts were provided for the same question keeping in mind that the meaning remains unchanged, and about an equal number of responses were collected from ChatGPT, Bard, and Aichatting.

C. Similarity Score Analysis

Once the responses were collected, Similarity Score Analysis was performed to analyze the resemblance between AI-AI, Human-Human, and AI-Human responses. It was performed by generating the embeddings for each response and evaluating the similarity scores.

1) BERT- Word Embedding Model: Word embedding involves language modeling and feature learning methods in NLP. These techniques map words or phrases from a natural language to vectors of real numbers. BERT (Bidirectional Encoder Representations from Transformers) [13,14] is a bidirectional pre-trained word embedding model used for various NLP-based tasks. It is used to extract features, particularly word and sentence embedding vectors from the whole text corpus. These embeddings help in capturing the semantic meaning of words by identifying the long-term dependencies by producing a word representation that is progressively implicated by words around them with the help of masking.

2) Similarity Score - Cosine Similarity: This technique evaluates how alike two embeddings are by measuring their similarity. The score is determined by the cosine angle formed between the two vectors being compared [15]. A greater cosine similarity value indicates that the vectors are closer, implying a higher degree of similarity between the two text corpora represented by the embeddings.

The cosine similarity between two vectors $\mathbf{u}$ and $\mathbf{v}$ is given by the formula:

$$\text{Cosine Similarity}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \cdot \|\mathbf{v}\|} \tag{1}$$

where $\cdot$ denotes the dot product and $\|\mathbf{u}\|$ represents the Euclidean norm (magnitude) of vector $\mathbf{u}$.

In this paper, the BERT embedding model has been used to generate the embeddings of all the responses. The cosine similarity score is calculated for all the possible human-human, human-AI, and AI-AI response pairs for respective questions and topics. Following this, corresponding mean scores are evaluated. The mean Human-Human similarity score indicates how similar are the two different human responses to a particular question. Similarly, the mean Human-AI similarity score indicates the similarity between human and AI responses, and, the mean AI-AI similarity score indicates the similarity of two AI-generated responses to a particular question.

D. Clustering

An unlabeled dataset containing human and AI responses for a particular question is taken, and BERT embeddings are generated on it, resulting in an unlabeled dataset consisting of vector embeddings.

Clustering is an unsupervised machine-learning algorithm. It is a method of distributing the data points into different clusters such that data points sharing similar features fall into the same group.

1) K-Means: K-means is one of the most popular clustering algorithms [16,17]. K-means can be summarized by following steps:

- Assume a value of K. Initialize the algorithm by assuming k random data points as centroids.
- Each data point in the dataset is assigned to the centroid at least a distance from it, thus forming k clusters. The distance measures used could be Euclidean or cosine distance.
- Calculate the variance and reassign the cluster centroids.
- Repeat 2 and 3 unless convergence is attained.

Determining an optimal number of clusters is a crucial step for K-means clustering. It is normally done by iterating the above process for different values of k.

2) Silhouette Score: The function within the scikit-learn library assesses the quality of formed clusters [18]. This score is determined by computing the mean silhouette coefficient for all samples in the dataset. The silhouette coefficient is derived from the average cluster cohesion (d1) and the average cluster separation (d2) for each data point. The silhouette coefficient for a sample is

$$\frac{d2 - d1}{\max(d1, d2)} \tag{2}$$

- When the value of the silhouette score is closer to +1, it means that the data points within the same cluster are highly similar.
- When the value of the silhouette score is near 0, it means that the data points within a similar cluster are poorly matched.
- When the value of the silhouette score is near -1, it means that the data points have been incorrectly placed in clusters.

A K-means model is trained on this dataset with k=2 as it is already known that two types of data points are present in the dataset. Cosine distance is used as the metric parameter for evaluating the distance between the data points. Further, the silhouette score of the model is evaluated.

E. Classification

The unlabelled dataset is analysed using various classification algorithms to classify the responses into two classes namely, human and AI using various features like vocabulary richness, spelling errors, readability score, etc.

Classification is a supervised machine-learning [19] method in which a classifier is trained on a labeled dataset and is subsequently used to predict the class of a new data point.

1) Types of Classifiers Used:

- **Logistic Regression**

  Logistic regression [20], is a classification model that uses the output generated by a linear regression function as an input followed by a sigmoid function that generates a probability for a given class. The class with the highest probability is the final output, therefore generating discrete outputs. It is an exhaustively used classification algorithm in the industry because of its impressive performance with linear classification as it makes use of a linear combination of input features to make decisions.

  $$P(Y=1) = \frac{1}{1 + e^{-(a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_n X_n)}} \tag{3}$$

  where:

  $P(Y = 1)$ is the probability of the dependent variable when it is a positive class

  e is the base of the natural logarithm,

  a_0 is the intercept term,

  $a_1, a_2, \ldots, a_n$ are the coefficients associated with features $X_1, X_2, \ldots, X_n$,

  $X_1, X_2, \ldots, X_n$ are the input features.

  While training a logistic regression model various hyperparameters need to be tuned to improve the performance of the model. Following are some important hyperparameters for logistic regression:

  - C: It is the inverse of regularization strength and hence prevents overfitting of the model. Its typical value is 1.
  - Penalty: It specifies the type of regularization to be applied (l1 or l2). The default value of penalty is ’l2’.
  - Solver: This hyperparameter specifies the optimization algorithm to be used for training the logistic regression model. The choice of solver can impact the convergence speed and efficiency of the model. The default value of the solver is ’lbfgs’. ’liblinear’, ’newton-cg’, ’sag’, ’saga’ are some other common values it can take.

- **Random Forest**

  This approach builds upon the bagging technique, incorporating both bagging and feature randomness to construct a diverse forest of decision trees [21,22]. The method involves generating multiple decision trees using distinct random subsets of the data and features. Predictions are then made by computing predictions for each decision tree and subsequently considering the most commonly occurring result [23]. The Random Forest classifier’s hyperparameters include:

  - max_depth: This parameter represents the longest path from the root node to a leaf node, limiting the depth required for the random forest to grow.
  - max_features: It indicates the maximum number of features provided to each tree in the random forest.
  - min_samples_split: This parameter sets the minimum required number of observations in any given node of a decision tree to initiate a split. The default value is 2.
  - min_samples_leaf: It specifies the minimum number of samples needed in a leaf node after a node has

been split. This parameter regulates tree growth by establishing a minimum sample criterion for terminal nodes.

- **XG Boost**
  XG Boost (Xtreme Gradient Boosting) is an implementation of gradient boosting that is specifically designed to be efficient and scalable [24]. Mathematically, it is a type of ensemble learning method that merges predictions from several weak models to generate a more robust prediction. The weak models in XGBoost are decision trees, which are trained using gradient boosting. This means that at each iteration, the algorithm fits a decision tree to the residuals of the previous iteration. The decision trees in XGBoost are trained using the following objective function:

$$\text{Objective Function} = \sum_{i=1}^{n} \text{loss}(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (4)$$

where:

n is the number of data points.

$\sum_{i=1}^{n}$ denotes the summation over all data points.

$\text{loss}(y_i, \hat{y}_i)$ represents the loss function, where y_i is the true label and $\hat{y}_i$ is the predicted label for the i-th data point.

K is the number of trees in the ensemble.

$\sum_{k=1}^{K}$ denotes the summation over all trees.

$\Omega(f_k)$ represents the regularization term for the k-th tree.

[25] The hyperparameters of XG Boost considered are:

  - max_depth: It tells the maximum depth of a tree and is used to control over-fitting. Increasing its value will more likely result to overfit.
  - min_child_weight: It tells the minimum sum of weights of all observations required in a child.

- **Support Vector Classifier**
  A machine learning model called Support Vector Classifier [26] is used for classification tasks. A set of training examples is given, where each of them belongs to one of the two classes. A support vector classifier can be used on them to separate the two classes maximally by finding a decision boundary. Mathematically, the decision boundary can be represented as

$$w_1 \cdot x_1 + w_2 \cdot x_2 + b = 0 \quad (5)$$

where:

w_1, w_2 : Weights or coefficients associated with features

x_1, x_2 : Input features or variables

b : Bias term, which shifts the hyperplane away from the origin

$= 0$: The equation represents the hyperplane where the decision boundary is defined.

The points that satisfy this equation form the decision boundary. To find the decision boundary that maximally separates the two classes, we can solve the following optimization problem:

$$\text{Minimize } \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\max(0, 1 - y_i(\mathbf{w}\cdot\mathbf{x}_i + b)) \quad (6)$$

where:

c denotes the regularization parameter that controls the trade-off between maximizing margin and minimizing error

y_i is the label of the i-th training example (either 1 or -1)

x_i is the corresponding feature vector

Once the optimization problem has been solved and the decision boundary has been found, we can use it to make predictions on new data points. Given a new feature vector x, we can predict the class of the corresponding example by computing the value of the linear function w^T * x+b and comparing it to 0. If the value is positive, we predict that the example belongs to one class; if it is negative, we predict that the example belongs to the other class.

2) Feature Extraction: From the responses collected through the survey, key features like linguistic, structural, and semantic attributes are extracted. be used to This methodology forms the foundation of the research to differentiate between human and AI-generated text.

Here's a concise overview of each feature and the methods employed to calculate them:

- **Vocabulary Richness** This feature quantifies the uniqueness of words within a given text. We calculate it by determining the ratio of unique words to the total number of words in the text.

$$\text{Vocabulary Richness} = \left(\frac{\text{Number of Unique Words}}{\text{Total Number of Words}}\right) \quad (7)$$

- **Spelling Errors** Spelling errors are indicative of the accuracy of the text. To determine the percentage of spelling errors in the text, we utilize the the PySpellChecker library.

$$\text{Spell Errors (\%)} = \left(\frac{\text{Words with Spelling Errors}}{\text{Total Words}}\right) \times 100 \quad (8)$$

- **Grammatical Errors** Grammatical correctness is another vital aspect of text quality. We assess the percentage of grammatical errors using the LanguageTool Package through the language-tool-python library.
- **Readability Score** This feature gauges the complexity and ease of interpretation of the text. We employ the textstat.flesch_kincaid_grade() function to calculate the Flesch-Kincaid grade level. This grade level estimation provides insights into the text's readability, where lower values correspond to easier-to-read text, and higher values suggest more complex text.
- **Label** A binary label assigned to each text, with "1" representing AI-generated responses/text and "0" denoting human-authored text.

TABLE I
Classifiers and Best Hyperparameters

Classifier	Best Hyperparameters
Logistic Regression	C=1, penalty='l1', solver='liblinear'

Random Forest	max_depth=10, max_features='log2',
	min_samples_split=2, min_samples_leaf=1

XGBoost	n_estimators=100, max_depth=3,
	min_child_weight=1

SVM (SVC)	C=1, kernel='rbf'

TABLE II
Mean Similarity Scores for different responses

Response Type	Mean-Similarity Score
Human-Human	0.6628739669711332
Human-AI	0.5005839041628254
AI-AI	0.7472267064040586

III. Results and Discussion

A. Similarity Score Analysis

It can be observed from Table II that if the similarity score of two or more responses is very high then it is likely to be the responses generated by AI. It means the responses of AI are consistent in terms of meaning and use the same set of words most of the time to answer a particular question.

B. K-means Clustering Analysis

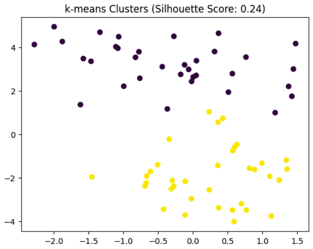

Fig. 2. Scatter-Plot for the K-Means Clustering

On observing the scatter plot of K-means clustering from Fig. 2 it can be said that the boundaries of the clusters created for the AI and human responses are not very much defined further justified by a silhouette score of 0.24. It can be because the clustering model was trained on AI and human responses to the same set of questions thus leading to responses of similar intent which the clustering model is unable to differentiate.

C. Classification Result

1) Logistic Regression: [27] This model yielded an accuracy of 0.833, a precision of 0.8125, recall of 0.8965, and an F1 score of 0.8524. While the performance is satisfactory, it lags behind other classifiers as per our study.

2) Random Forest: With the tuned hypermeters, the model achieved an accuracy of 0.8518, a precision of 0.85, a recall of 0.88, and an F1 score of 0.86. Notably, the accuracy improved compared to Logistic Regression, making an overall enhancement in the results.

3) XG Boost: The combination of set ideal hypermeters delivered an accuracy of 0.87, a precision of 0.866, a recall of 0.896, and an F1 score of 0.881. Significantly, XGBoost exhibited even better accuracy than Random Forest, demonstrating superior performance across all other metrics.

4) Support Vector Machine: SVM resulted with an accuracy of 0.87, a precision of 0.84, a recall of 0.93, and an F1 score of 0.88. While the accuracy remains consistent with XGBoost, the high recall rate indicates superior classification performance, making SVM (SVC) a compelling choice for the task at hand.

5) Comparision: In our evaluation of the classifiers, the SVM (SVC) model, with its optimized hyperparameters, emerged as the top performer, showcasing an exceptional recall rate of 0.93, making it highly proficient at accurately identifying AI-generated text. Moreover, it maintained a well-balanced overall performance with a precision of 0.84 and an F1 score of 0.88, along with an accuracy of 0.87, which was on par with XGBoost. On the other end of the spectrum, Logistic Regression delivered the least competitive results, with lowest precision and accuracy of 0.8125 and 0.833 respectively.

TABLE III
Classifier Metrics

Classifier	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.833	0.8125	0.8965	0.8524
Random Forest	0.8518	0.85	0.88	0.86
XGBoost	0.87	0.866	0.896	0.881
SVM (SVC)	0.87	0.84	0.93	0.88

6) Correlation curve: From the correlation curve in Fig. 3, positive correlation indicates that increasing the values of these features is associated with a higher likelihood of the label being 'non-human.' In contrast, a negative correlation suggests that a decrease in the feature value corresponds to a higher probability of the text being classified as 'human.' 'Grammar_Error' emerged as the most prominent factor in identifying human text from AI-generated content, with a substantial negative correlation coefficient of -0.4, indicating that humans tend to make more grammatical errors, while AI-generated text exhibits better grammatical accuracy.

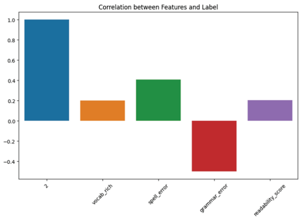

Fig. 3. Correlation curve between Features and Labels

7) Receiver Operating Characteristic Area (ROC - AUC) : The ROC curve [28] (Fig. 4) visualizes the trade-off between True Positive Rate and False Positive Rate.

The AUC of '0.86' indicates that the model can distinguish between positive and negative cases with a high degree of accuracy. It is also able to rank positive cases higher than negative cases with a high degree of accuracy

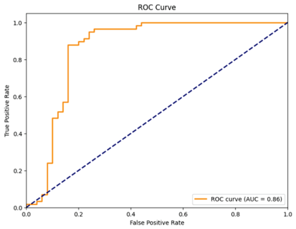

Fig. 4. ROC-AUC Curve

8) Precision Recall Curve (PR): The area under the Precision-Recall Curve tells the summary measure that quantifies the model's ability to correctly classify positive instances while maintaining high precision.

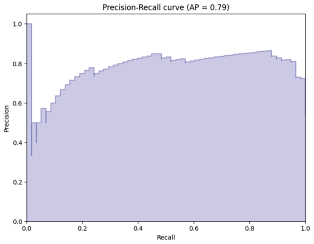

Fig. 5. Precision-Recall Curve

Average Precision (AP) of 0.79 indicates that the model performs reasonably well in terms of precision and recall. Generally, higher values of AP are desirable, as they signify a model that can effectively identify positive instances while maintaining a low rate of false positives.

References

[1] Neha Saini, "Artificial Intelligence and its Applications" *International Journal of Science & Engineering Development Research*, vol. 8, issue 4, pp 356-360, April 2023.

[2] Mijwil, Maad. (2015). History of Artificial Intelligence. 3. 1-8. 10.13140/RG.2.2.16418.15046.

[3] Sah, Shagan. (2020). Machine Learning: A Review of Learning Types.

[4] Khurana, Diksha & Koli, Aditya & Khatter, Kiran & Singh, Sukhdev, "Natural Language Processing: State of The Art, Current Trends and Challenges" in *Multimedia Tools and Applications*, pp 3713-3744, July 2022.

[5] Tiwari, Rudra. "The Impact of AI and Machine Learning on Job Displacement and Employment Opportunities" in *International Journal of Engineering Technologies and Management Research*, vol, 7, issue 1, January 2023.

[6] Feuerriegel, Stefan and Hartmann, Jochen and Janiesch, Christian and Zschech, Patrick. "Generative AI", *Business & Information Systems Engineering*, May 2023.

[7] R. A. Partadiredja, C. E. Serrano and D. Ljubenkov, "AI or Human: The Socio-ethical Implications of AI-Generated Media Content" in *13th CMI Conference on Cybersecurity and Privacy (CMI) - Digital Transformation - Potentials and Challenges (51275)*, Copenhagen, Denmark, 2020, pp. 1-6.

[8] Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J "Ethical Considerations of Using ChatGPT in Health Care" in *J Med Internet Res* 2023.

[9] Illia, L., Colleoni, E., Zyglidopoulos, S. "Ethical Implications of Text Generation in the Age of Artificial Intelligence", in *Business Ethics, the Environment & Responsibility* 2023; 32: 201–210.

[10] Wang L, Zhou L, Yang W, Yu R., "Deepfakes: A new threat to image fabrication in scientific publications?", *Patterns (N Y)*, May 2022.

[11] Kai-Qing Zhou and Hatem Nabus, "The Ethical Implications of DALL-E: Opportunities and Challenges" in *Mesopotamian Journal of Computer Science*, vol. 2023, pp. 17–23, Feb. 2023.

[12] A.M. Turing, "Computing Machinery and Intelligence" in *Oxford University Press* Vol. 59, No. 236, pp 433-460, Oct 1950.

[13] Tanaka, H., Shinnou, H., Cao, R., Bai, J., Ma, W. "Document Classification by Word Embeddings of BERT" in Nguyen, LM., Phan, XH., Hasida, K., Tojo, S. in *Computational Linguistics. PACLING 2019. Communications in Computer and Information Science*, vol 1215. Springer, Singapore, 2019, pp 145-154.

[14] R. K. Kaliyar, "A Multi-layer Bidirectional Transformer Encoder for Pre-trained Word Embedding: A Survey of BERT" in *10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, 2020, pp. 336-340.

[15] A. R. Lahitani, A. E. Permanasari and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment" in *4th International Conference on Cyber and IT Service Management*, Bandung, Indonesia, 2016, pp. 1-6.

[16] K. P. Sinaga and M. -S. Yang, "Unsupervised K-Means Clustering Algorithm," in *IEEE Access*, vol. 8, pp. 80716-80727, 2020.

[17] K. Bindra and A. Mishra, "A detailed study of clustering algorithms" in *6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India, 2017, pp. 371-376.

[18] K. R. Shahapure and C. Nicholas, "Cluster Quality Analysis Using Silhouette Score" in *IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, Sydney, NSW, Australia, 2020, pp. 747-748.

[19] S. Dridi, *"Supervised Learning - A Systematic Literature Review"*, 04-Apr-2022. [Online]. Available: osf.io/tysr4.

[20] Z. A. A. A. Bairmani and A. A. Ismael, "Using Logistic Regression Model to Study the Most Important Factors Which Affects Diabetes for The Elderly in The City of Hilla in *Journal of Physics*, vol. 1818, no. 1, p. 012016, Mar. 2021.

[21] Mienye, Domor, Sun, Yanxia. (2022). "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects" in *IEEE Access*, vol. 10, pp. 99129-99149, 2022.

[22] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning" in *Stata Journal*, vol. 20, no. 1, pp. 3–29, Mar. 2020.

[23] Sharoon Saxena (2023, Aug 25) *A Beginner's Guide to Random Forest Hyperparamter Tuning* [Online]. Available: https://www.analyticsvidhya.com/blog/2020/03/beginners-guide-random-forest-hyperparameter-tuning/

[24] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A Comparative analysis of XGBooSt" in *Cornell University*, Nov. 2019.

[25] Prashant Banerjee (2020) *A Guide on XGBoost hyperparameters tuning* [Online]. Available: https://www.kaggle.com/code/prashant111/a-guide-on-xgboost-hyperparameters-tuning

[26] J. Zhao, "The development and application of Support Vector Machine" in *Journal of Physics*, vol. 1748, no. 5, p. 052006, Jan. 2021.

[27] Ž. Vujović, "Classification model evaluation metrics", in *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, Jan. 2021.

[28] Davis J., Goadrich M."The relationship between Precision-Recall and ROC curves" in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006, pp. 233-240.