Machine Learning Model for Distinguishing Human and ChatGPT Text

Likith Kumar Dundigalla
School of Information
University of Arizona
Tucson, AZ 85721
likithkumard@arizona.edu

Abstract

In this project, three distinct approaches are explored to distinguish between human-generated and AI-generated text. The existing approach involves generating TD-IDF vectors and comparing their accuracies and F1 scores across multiple models. The prediction is made based on the model that achieves the highest score.The N-grams approach focuses on transforming text data into n-gram features using tokenization and CountVectorizer. This allows the representation of word sequences as features for classification using Multinomial Naive Bayes. Additionally, the Word Embeddings approach leverages Word2Vec models to embed text data into numerical vectors, capturing semantic information. It utilizes Logistic Regression for classification based on these embedded representations. These approaches were designed to encapsulate different aspects of text representation and feature extraction. They aim to capture nuanced patterns and semantic information within the text, improving classification accuracy compared to the baseline online approach. The exploration of N-grams and Word Embeddings provides a more comprehensive understanding of text data and its underlying structures. This exploration potentially offers enhanced performance in distinguishing between human and AI-generated text.

1 Introduction

Have you ever wondered how to tell if a text was written by a human or an AI? Do you think you could tell the difference? 'Can you tell me about the history of the Kohinoor (Koh-i-Noor) Diamond?'. Try to guess who generated this question: A human or an AI. With the advances in natural language generation, it is becoming harder and harder to tell the difference. In this project, we will build a machine-learning models that will tell if a human or ChatGPT generated the text. Sounds interesting, right? Let's get started.

The project revolves around the task of discerning between human-generated and AI-generated text. The primary objective is to create machine learning model capable of accurately classifying text into these two categories.

To achieve this, the project employs various Natural Language Processing (NLP) techniques and machine learning algorithms. Three distinct methods are proposed and implemented:

1.1 Online Approach

This approach focuses on extracting features from the provided text dataset and training multiple classifiers, such as Logistic Regression, Support Vector Machines (SVM), Decision Trees, and others.

The method involves transforming the text data, extracting relevant features, training classifiers, and evaluating their accuracy using confusion matrices. This approach utilizes techniques like TF-IDF vectorization and various classification algorithms to distinguish between human and AI-generated text.

1.2 Ngrams Approach

This approach employs Ngrams, which are continuous sequences of words, to tokenize and create features from the text data. By utilizing CountVectorizer to generate Ngrams features and employing a Multinomial Naive Bayes classifier, this method aims to classify text into human-generated or AI-generated categories.

1.3 Word Embeddings Approach

Utilizing Word2Vec, a technique for learning word embeddings, this approach transforms text data into numerical vectors to capture semantic similarities between words. The Word2Vec model converts text into vectors, which are then used to train a Logistic Regression classifier to differentiate between human-generated and AI-generated text.

Each method undergoes a similar process of data transformation, feature extraction, model training, and evaluation. The evaluation metrics primarily include accuracy scores and confusion matrices to assess the performance of the models.

These methods are implemented with the goal of demonstrating the effectiveness of different NLP techniques and machine learning algorithms in correctly categorizing text as either human-generated or AI-generated.

2 Related Work

[1] How to Build a Machine Learning Model to Distinguish If It's Human or ChatGPT?

Link for the reference:`https://www.analyticsvidhya.com/blog/2023/04/how-to-build-a-machine-learning-model-to-distinguish-if-its-human-or-chatgpt/#Implementation`
Amrutha K(Analytics Vidhya)
The objective entails constructing a machine-learning model capable of distinguishing between human-generated and ChatGPT-produced text across various genres, including questions, essays, stories, jokes, code, and more. The aim is to develop a versatile classifier capable of handling diverse textual formats.

[2] Check Me If You Can: Detecting ChatGPT-Generated Academic Writing using CheckGPT

Link for the reference: `https://arxiv.org/abs/2306.05524`
Zeyan Liu, Zijun Yao, Fengjun Li, Bo Luo
This paper investigates detecting LLM-generated academic writing, offering a dataset, evidence, and algorithms. It introduces GPABenchmark, a dataset with 600,000 human-written, GPT-generated, completed, and polished abstracts in various research fields. Existing GPT detectors perform poorly on GPABenchmark, especially for polished text. A user study confirms the challenge for humans to identify GPT-generated abstracts. CheckGPT, a novel LLM-content detector, achieves 98% to 99% accuracy for discipline-specific and unified detectors. It exhibits 90% accuracy in new domains without tuning and reaches 98% accuracy with domain-specific tuning. The paper also demonstrates explainability insights from CheckGPT regarding key behaviors in LLM-generated texts.

[3] AI vs. Human – Differentiation Analysis of Scientific Content Generation

Link for the reference: `https://arxiv.org/abs/2301.10416`
Yongqiang Ma, Jiawei Liu, Fan Yi, Qikai Cheng, Yong Huang, Wei Lu, Xiaozhong Liu
This study delves into the nuanced differences between AI-generated scientific content and human-written text, particularly in the context of scientific writing assistance. We present a feature-based

framework encompassing syntax, semantics, and pragmatics to discern these distinctions via human evaluation. Analyzing writing style, coherence, consistency, and argument logistics, we compare two content types and employ various detection methods to explore gaps between AI-generated and human-written scientific text. Our findings reveal that although AI demonstrates potential in accuracy, discrepancies persist in depth and overall quality, often manifesting in factual errors. Notably, a distinct "writing style" gap is observed. We summarize model-agnostic and distribution-agnostic features for detection across domains. This research aims to steer AI model optimization for improved content quality, addressing pertinent ethical and security concerns.

3 Procedure

3.1 Online Approach

The project employs a diverse set of machine learning models to distinguish between human and AI-generated text. The process begins with the extraction and transformation of text data from the provided dataset. The models used include Logistic Regression, Support Vector Machines (SVM), Decision Trees, K-Nearest Neighbors (KNN), Random Forest, Extra Trees, AdaBoost, Bagging, and Gradient Boosting classifiers.

The text data is vectorized using the TF-IDF (Term Frequency-Inverse Document Frequency) technique, which represents the importance of words in a document. Each model is then trained on the TF-IDF vectorized data to learn the patterns distinguishing between human and AI-generated text. After training, the models make predictions on test data, and their accuracy is evaluated using confusion matrices. This comprehensive approach aims to leverage the strengths of various classifiers and TF-IDF vectorization to achieve accurate classification of text into human or AI-generated categories. The entire process involves extracting meaningful features from text, training multiple classifiers, and evaluating their performance in distinguishing between the two text sources.

3.1.1 TF-IDF Explanation

TF-IDF (Term Frequency-Inverse Document Frequency) is a technique used in natural language processing to evaluate the importance of a word in a document relative to a collection of documents. It consists of two components:

Term Frequency (TF): Measures the frequency of a term (word) within a document. Mathematically represented as:

$$\textbf{TF(t, d)} = \frac{\text{Number of times term } word \text{ appears in document}}{\text{Total number of words in document}}$$

Inverse Document Frequency (IDF): Measures the rarity of a term across all documents in a dataset. Mathematically represented as:

$$\textbf{IDF(t, D)} = \log\left(\frac{\text{Total number of documents in the dataset}}{\text{Number of documents containing the word}}\right)$$

The overall TF-IDF score for a term in a document combines these two components by multiplying the TF and IDF scores together:

$$\textbf{TF-IDF(t, d, D)} = TF(t, d) \times IDF(t, D)$$

This score represents the importance of a term in a specific document within a collection of documents. Higher TF-IDF scores indicate that a term is more important or relevant to the document.

3.1.2 TF-IDF in Online Approach

In the context of the Online Approach, TF-IDF vectorization is used to convert the text data into numerical vectors. These vectors represent the importance of words within each document relative to the entire dataset. This process helps in transforming the textual data into a format suitable for training machine learning models like Logistic Regression, Support Vector Machines, and others, allowing these models to learn from the TF-IDF weighted features and effectively distinguish between human and AI-generated text.

3.2 Extra Trees Classifier(ETC)

ETC is an ensemble learning method based on decision tree classifiers. It creates a forest of randomized decision trees and aggregates their predictions. "Extra" in Extra Trees refers to the fact that it chooses random thresholds for each feature, hence adding more randomness compared to regular decision trees or random forests. This randomness often helps in reducing overfitting and can be beneficial when dealing with high-dimensional data like text.
The ETC model here is trained using the TF-IDF transformed training data.

3.3 Ngrams Approach

N-grams play a crucial role in text classification tasks by capturing language patterns, differentiating features, and improving model understanding. The choice between unigrams and bigrams affects the granularity of the features extracted.

Unigrams (1-grams): These are single words considered individually.
Bigrams (2-grams): These consist of pairs of adjacent words.

In our approach, `CountVectorizer` is configured with `ngram_range=(1, 2)`, enabling the extraction of both unigrams and bigrams, covering both individual words and pairs of words in the text data for classification.

3.3.1 Ngrams in Ngrams Approach

The Ngrams approach utilizes the Multinomial Naive Bayes (MNB) classifier to predict the category (human-generated or AI-generated) of a given text based on the frequency distribution of Ngrams. The mathematics behind this classifier involves probability estimation using Bayes' theorem and assumes conditional independence between features.

Bayes' Theorem:

$$\mathrm{P}(c \mid x) = \frac{P(x \mid c) \cdot P(c)}{P(x)}$$

Where:
$P(c \mid x)$ is the posterior probability of class c given the feature vector x.
$P(x \mid c)$ is the likelihood, the probability of observing the feature vector x given the class.
$P(c)$ is the prior probability of class c.
$P(x)$ is the probability of observing the feature vector x.

In the case of the Ngrams approach, the MNB classifier estimates $P(x \mid c)$, the likelihood of observing a particular Ngram sequence given a class. This calculation involves the frequency distribution of Ngrams in the training data for each class (human-generated or AI-generated).

3.4 Word Embeddings Approach

In the Word Embeddings approach, the primary model utilized is Word2Vec. This technique is designed to represent words as high-dimensional vectors in a continuous space where the relationships between words are captured based on their context in the text corpus.

3.4.1 Word2Vec in Word Embeddings Approach

Word2Vec utilizes two primary architectures: Continuous Bag of Words (CBOW) and Skip-gram.

Continuous Bag of Words (CBOW): CBOW predicts a target word from its neighboring context words. Mathematically, CBOW maximizes the probability of predicting the target word given its context:

$$\max \frac{1}{T} \sum_{t=1}^{T} \log p(w_t \mid \text{context}(w_t))$$

Skip-gram: In contrast, Skip-gram predicts context words using a target word. It maximizes the likelihood of predicting the context words around a target word:

$$\max \frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} \mid w_t)$$

During training, both CBOW and Skip-gram use a softmax function to compute the probability distribution over the vocabulary for predicting the target or context words.

3.4.2 Logistic Regression in Word Embeddings Approach

Once Word2Vec learns the word embeddings, Logistic Regression is employed as a classifier. The logistic function computes the probability of a text belonging to a category using the word embeddings:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(b + \sum_{i=1}^{n} w_i \cdot x_i)}}$$

Where: $P(y = 1 \mid x)$ represents the probability of the text being in the "AI-generated" category. b is the bias term. w_i are the weights. x_i are the word embeddings/features. n is the number of features (dimension of word embeddings).

The Word2Vec model captures semantic relationships between words based on their context in the text corpus. These word embeddings are then utilized by Logistic Regression for accurate text classification based on the learned relationships.

4 Evaluation

4.1 Online Approach Evaluation

The Online Approach initially focused on extracting features and training various classifiers, including Logistic Regression, Support Vector Machines (SVM), Decision Trees, and others. The TF-IDF vectorization technique was employed to represent text data numerically. The models were evaluated primarily based on their accuracy scores and confusion matrices.

4.1.1 Model Performance

Each classifier exhibited varied performance in distinguishing between human-generated and AI-generated text. Extra Tree Classifier demonstrated considerable accuracy, achieving notable discrimination between the two categories. However, KNeighbors and Decision Trees Classifiers showed relatively lower accuracy scores.

Table 1: Model Evaluation Scores

Model	Accuracy	F1 Score
Logistic Regression	0.8311	0.8320
Support Vector Machine	0.8152	0.8171
Multinomial Naive Bayes	0.7572	0.7611
Decision Tree Classifier	0.7006	0.6951
KNeighbors Classifier	0.6185	0.5193
Random Forest Classifier	0.8250	0.8290
Extra Trees Classifier	0.8350	0.8375
AdaBoost Classifier	0.7609	0.7617
Bagging Classifier	0.8027	0.8083
GradientBoosting Classifier	0.7484	0.7547

*Average of 5 executions is taken.

4.1.2 Confusion Matrices

The confusion matrices provided a detailed understanding of the models' performance, indicating the number of true positives, true negatives, false positives, and false negatives. These matrices

depicted the classifiers' ability to correctly classify human and AI-generated text, offering insights into misclassifications and error patterns.

4.2 Ngrams Approach Evaluation

The Ngrams Approach aimed to tokenize text data and leverage Multinomial Naive Bayes (MNB) for classification based on Ngrams' frequency distributions.

4.2.1 Feature Extraction

The utilization of `CountVectorizer` with `ngram_range=(1, 2)` enabled the extraction of both unigrams and bigrams. This approach enhanced the granularity of features by capturing sequences of words, potentially improving the models' ability to discern subtle linguistic patterns.

4.2.2 Classification Performance

The evaluation of the Ngrams Approach highlighted its effectiveness in categorizing text. The Multinomial Naive Bayes classifier demonstrated reasonably good performance, indicating promising capabilities in distinguishing between human and AI-generated text based on Ngrams' frequency distribution.

4.3 Word Embeddings Approach Evaluation

The Word Embeddings Approach utilized Word2Vec to transform text into numerical vectors, capturing semantic relationships between words for classification via Logistic Regression.

4.3.1 Semantic Representations

Word2Vec facilitated the creation of word embeddings that encapsulated semantic information. The embeddings were utilized as features to train the Logistic Regression classifier, enabling it to capture nuanced semantic relationships between words.

4.3.2 Classification Accuracy

The Word Embeddings Approach showed promising accuracy in discerning between human-generated and AI-generated text. The model's performance highlighted the significance of leveraging semantic context encoded within word embeddings for accurate text classification.

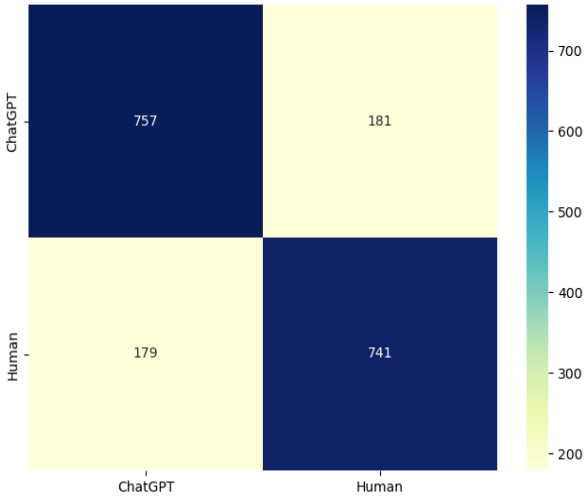
5 Results

5.1 Online Approach

The Online Approach showcased commendable performance with an accuracy of approximately 80%. Among the ensemble of classifiers used, Extra Trees Classifier (ETC) emerged as the top performer within this approach, achieving the highest accuracy among the models at 80%. The confusion matrices provided insightful details about the distribution of accurate predictions across both human and AI-generated text categories.

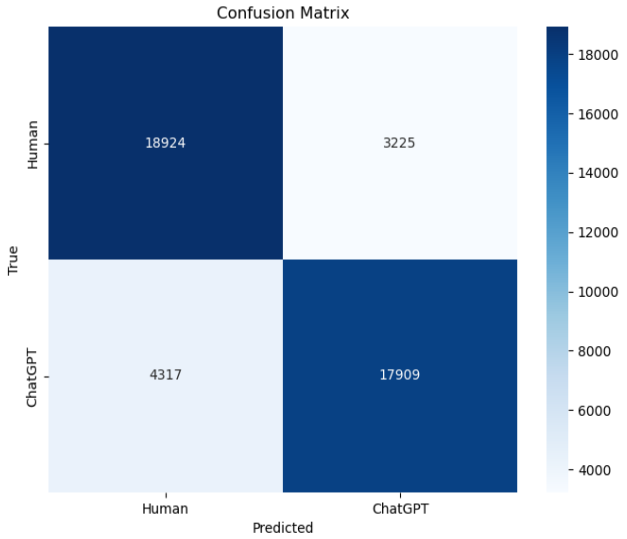
Table 2: Category Comparison with Text

Category	ID	Text	Predicted Category
chatgpt	21	The boo	chatgpt
chatgpt	22	As a second-year student of PESU's EEE progra	chatgpt
human	23	How can I convince my family for marriage?	human
chatgpt	24	Will Narendra Modi be elected as the Prime Min...	chatgpt
human	25	What is actual situation of petroleum engineer...	chatgpt
human	26	He will also talk about new ways NATO and the ...	chatgpt
human	27	I want to build one marriage hall in my land, ...	human
chatgpt	28	What is the root of misogyny?	human
human	29	Cavanaugh also presided over the construction ...	human
chatgpt	30	What is causing my dog to shake excessively?	chatgpt



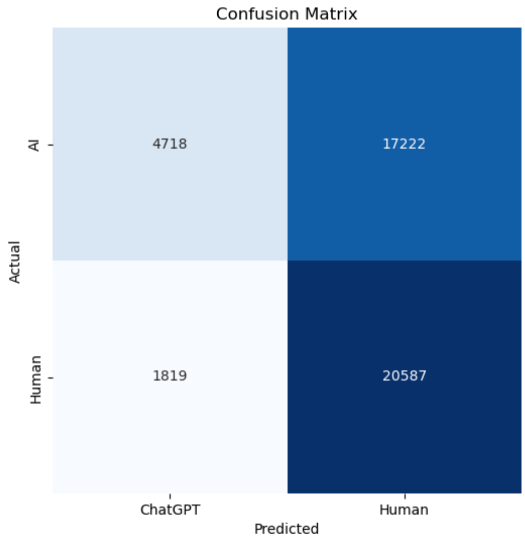
5.2 Ngrams Approach

Utilizing Ngrams (both unigrams and bigrams) and Multinomial Naive Bayes, the Ngrams Approach aimed to classify text based on the frequency distribution of Ngrams. The accuracy achieved by this approach was notably higher compared to the Online Approach, hovering around 83%. The Ngrams Approach revealed intriguing patterns in the confusion matrices, indicating specific strengths and weaknesses in classifying certain types of text.



5.3 Word Embeddings Approach

The Word Embeddings Approach, utilizing Word2Vec models and Logistic Regression, attained an accuracy of approximately 57%. Although the accuracy was comparatively lower than the other approaches, it provided valuable insights into semantic relationships between words, enhancing the understanding of text distinctions between human and AI-generated sources.



6 Conclusion

The evaluation of the three distinct approaches illustrated varying levels of accuracy in distinguishing between human and AI-generated text. The Ngrams Approach surpassed the other methods, demon-

Table 3: Accuracies Comparision of 3 Models

Model	Records	Accuracy
Online Approach	4000	80
N-grams Approach	125000	83
Word Embeddings	125000	57

*Average of 5 executions is taken.

strating the highest accuracy at 83%. The Online Approach followed closely, showcasing competitive accuracy at 80%, particularly with the Extra Trees Classifier. While the Word Embeddings Approach lagged behind in accuracy, its emphasis on capturing semantic nuances within text offered supplementary insights into classification tasks. The results underscore the significance of different feature extraction techniques and classification algorithms in text classification tasks, paving the way for further exploration and refinement of these approaches.

7 References

https://chat.openai.com/share/1954395d-66ba-4d60-8bb7-68aec664fe23
https://www.turing.com/kb/guide-on-word-embeddings-in-nlp

[1] Online Approach

https://www.analyticsvidhya.com/blog/2023/04/how-to-build-a-machine-learning-model-to-distinguish-if-its-human-or-chatgpt/#Implementation
https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/
https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-selection/
https://monkeylearn.com/blog/what-is-a-classifier/

[2] N-grams Approach

https://chat.openai.com/share/46d12112-3991-4539-9a17-e405f5c2ffff
https://www.geeksforgeeks.org/n-gram-language-modelling-with-nltk/
https://www.datacamp.com/blog/what-is-tokenization
https://www.analyticsvidhya.com/blog/2021/09/what-are-n-grams-and-how-to-implement-them-in-python/#:~:text=N%2Dgrams%20are%20continuous%20sequences,(Natural%20Language%20Processing)%20tasks.

[3] Word Embedding Approach

https://chat.openai.com/share/9344d764-da60-460b-932c-3766d41db4d1
https://www.geeksforgeeks.org/word-embeddings-in-nlp/
https://www.tensorflow.org/text/tutorials/word2vec
https://www.kaggle.com/code/kstathou/word-embeddings-logistic-regression