AI-Generated Content Detection

Gillian Tatreau
Department of Data Science, Bellevue University
DSC680: Applied Data Science
Amirfarrokh Iranitalab
2 March 2024

Business Problem
With the rise in popularity of AI-based tools and increased access to these resources, AI-generated content is becoming increasingly more common. It is found throughout the internet and has even made its way into academia. Content generated by AI is typically generated around relevancy- how closely related the words are to a prompt. This does not mean that this content is always accurate or correct. Being able to detect if a body of text was generated by AI would be beneficial for writers, publishers, educators, and students. Using predictive machine learning algorithms, bodies of text will be classified as either human or AI-generated. In doing so, it will be determined whether it is possible to differentiate between AI-generated content and human-generated text.

Background/History
Generative AI has been actively researched since the 1960s, with the advent of the world's first chatbot, ELIZA (White). The introduction of various consumer-targeted AI model services in the past couple of years has led to a meteoric rise in popularity of AI-generated text. Services such as ChatGPT can produce text in a variety of topics with an ever-expanding range of tones. The prevalence of AI-generated text and ease with which it is generated has allowed for a commodification of this technology, which has led to wide-spread adoption and even misuse.

Data Explanation
The dataset was found on Kaggle. It is composed of approximately 500,000 essays (Gerami).

Data Dictionary:
- text: body of text/paper (str)
- generated: indicates if text is human (0) or AI-generated (1) (binary)

Rows that did not contain a period were removed in order to calculate the number of sentences in each text and the average number of words per sentence. 293 human-generated texts and 79 AI-generated texts were removed from this step. Some of these entries contained only the newline tags '\n\n' and no actual text.

Sentence Length Models Data Preparation
The variables, number_sentences and words_per_sentence, were created. Data was standardized using the StandardScaler() function.

TF-IDF Models Data Preparation
Newline tags, punctuation marks, and stop words were removed from the text data.

Methods
The data was examined using graphical and statistical methods. Detection of AI-generated content represents a binary classification problem in this dataset, and thus classification machine learning algorithms were undertaken based off two different methodologies: sentence length and TF-IDF (term frequency inverse document frequency) values. The machine learning algorithms trained include logistic regression models, random forest classifiers, naïve Bayes classifiers, and support vector classification models. The models are evaluated by their accuracy, recall, and precision scores, as well as by a confusion matrix.

Analysis

The text data is approximately two-thirds human-generated text and one third AI-generated text.

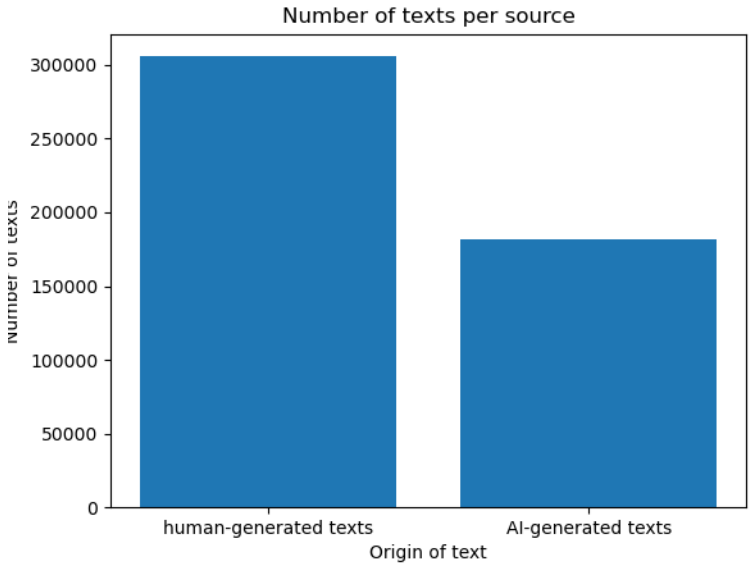

Figure 1. Distribution of origin sources within the dataset.

The distributions of both number_sentences and words_per_sentence appear almost normally distributed, with outliers present on the upper extreme of both variables. Outliers are not removed in an effort to preserve data that relates to the complexity and variability of the data. Run-on sentences could explain the presence of large outliers in the words_per_sentence variable. Some texts are also longer than others, which would explain the variation in number_sentences.

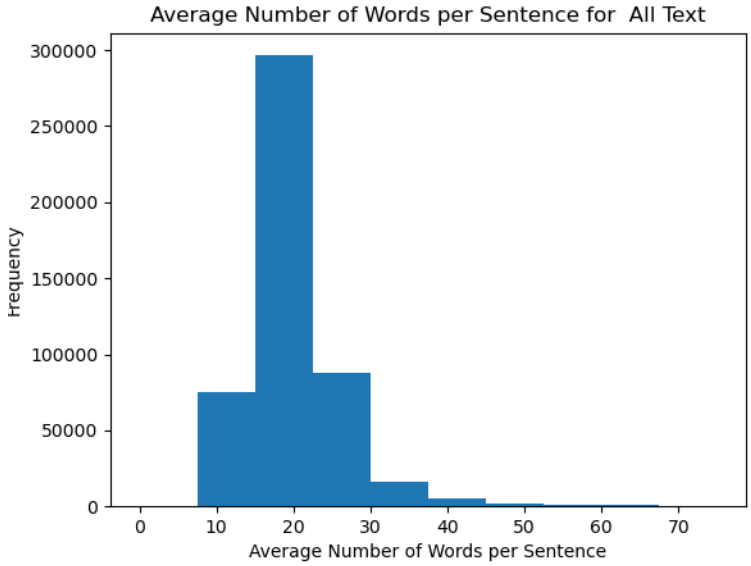

Figure 2. Distribution of words_per_sentence variable.

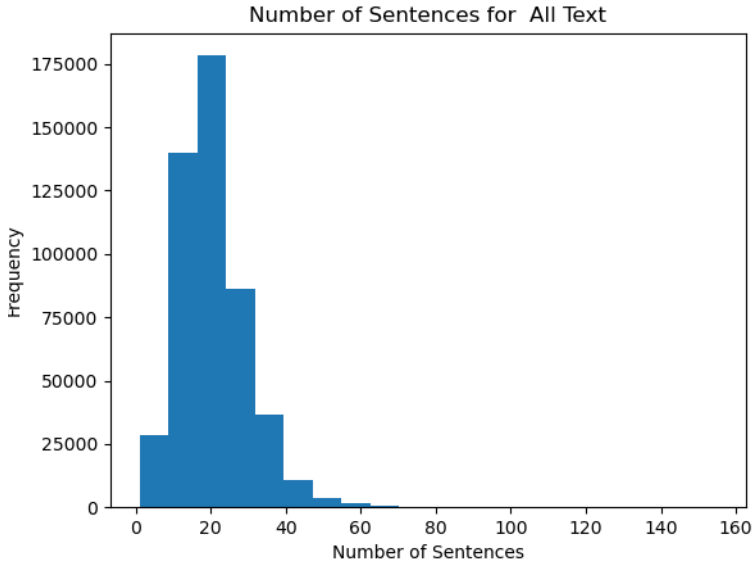

Figure 3. Distribution of number_sentences variable.

When differentiating between the two sources of text, it becomes clear that the human-generated text experiences far more variability. The AI-generated text experiences a positive kurtosis, as demonstrated by the sharp peaks in both plots. The human-generated text is far less strongly peaked and might possibly experience a negative kurtosis. Many of the outliers can then seemingly be attributed to the human-generated text and the variability inherent in human writing.

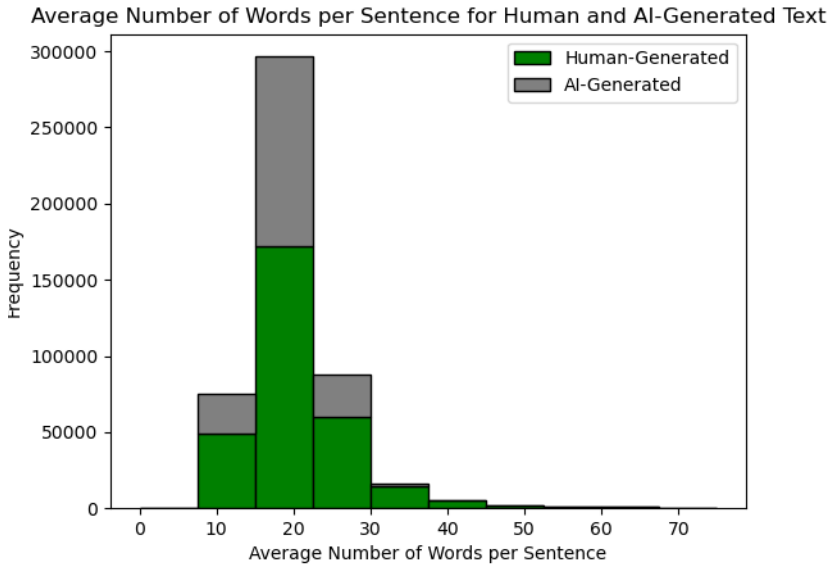

Figure 4. Distribution of words_per_sentence, with differentiation present between the sources of the text: green is human-generated text and grey is AI-generated text.

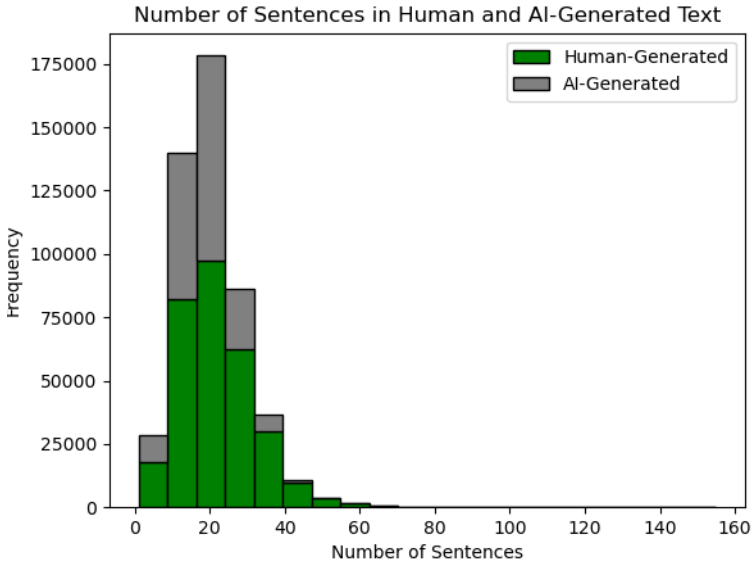

Figure 5. Distribution of number_sentences variable, with differentiation present between the sources of the text: green is human-generated text and grey is AI-generated text.

Data was randomly split into testing and training subsets, where the training data represented 80% of the total data and the testing data represented the remaining 20%.

Models Based on Sentence Length and Variability

Basic models of each machine learning algorithm were built and trained on the training data. Accuracy was calculated on both the training and testing data and compared to a baseline dummy classifier. Of the four models trained, only the support vector classifier and the random forest classifier models outperformed the dummy model. Of these two, the random forest far surpassed any of the other models for methodology (see appendix 1 for evaluation metrics of all initial models).

Hyperparameter tuning of the random forest classifier was performed using the RandomizedSearchCV() function. After hyperparameter tuning was performed, the best random forest model achieved adequate success. The accuracy score on the testing data was 0.7022. The precision score was 0.6141. The recall score was 0.5481.

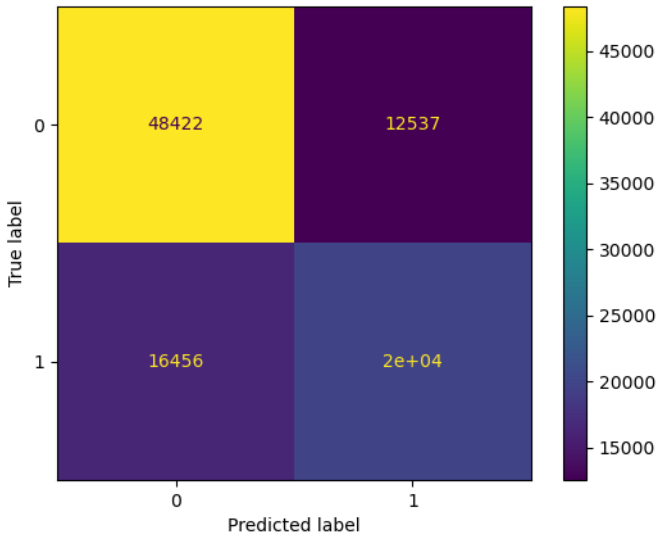

Figure 6. Confusion matrix of random forest model based on sentence length and variability.

This model struggles with both type I and type II errors, although it seems to struggle slightly more in correctly identifying AI-generated text, as there are more incorrectly predicted AI-generated texts than incorrectly predicted human-generated texts.

Models Based on TF-IDF

The data is slightly more complex for these models and thus take longer to train. To combat this, the initial models were trained on a subset of the split data, where only the first 80,000 rows of training data were used, and the first 20,000 rows of testing data were used. The three initial models trained in this methodology were the random forest classifier, the naïve bayes classifier, and the support vector classifier. All three initial models outperformed the dummy classifier. Their performance was so high that the precision and recall scores were calculated for these initial models to determine the best model for this treatment of the data. While the performance of all three models was similar, the random forest classifier again outperformed the other two in all three evaluation metrics (see appendix 2 for evaluation metrics of all baseline models).

The random forest classifier is then trained on the unabbreviated training data and the unabbreviated testing data is used to evaluate the model's performance. The accuracy score was 0.99857. The precision score was 0.99961. The recall score was 0.99657.

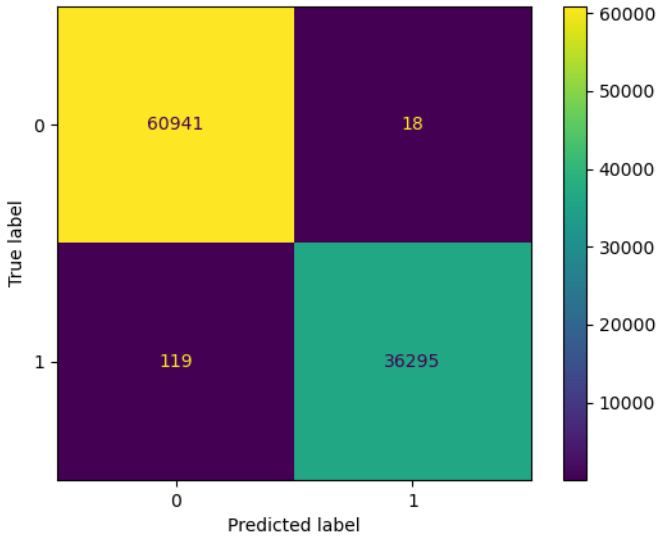

Figure 7. Confusion matrix for random forest model based on TF-IDF.

This model greatly outperforms any other built on this data.

Conclusion

While AI-generative technology has improved rapidly over the past decade, it is still in its infancy; therefore, it is reasonable to conclude that it is possible to differentiate AI-generated text from human-generated text. Some simple differences include the variability in the sentence structure and length. Human-generated text varies widely with the author's writing style while AI-generated text is more cohesive.

The random forest model trained using the TF-IDF data is a high-performing model. Even across various evaluation metrics, this model performs reliably and has few shortcomings. The random forest model trained on the sentence length data performs adequately, but the large amount of misidentified data and especially the amount of AI-text incorrectly labelled as human-generated calls into question the applicability of this model.

Assumptions

The first assumption is that all the text was written entirely in English, as only English stop words are removed. The second assumption made is that the essays were attained from publicly available sources or were donated to the dataset with the author's consent.

Challenges

One of the challenges with this data is the sheer size of the dataset, which made training the models time intensive. Another challenge was the potential of overfitting that is common with decision tree and random forest classifier models. The first random forest model trained on the abbreviated TF-IDF data was overfitted, as demonstrated by the perfect training accuracy score (see appendix 2 for the evaluation metrics). While the model trained on all the training data was not so obviously overfitted, it calls into question the potential applicability of this model without first seeing how well it behaves on additional data.

Future Uses
A deployed model that can detect AI-generated content could have many uses. It can be used as a grading tool used by schools to unsure that students are turning in their own work. It can also be used in resume and applicant screening. A model of this nature could also be used to vet articles before articles are posted online or published on news sites, blogs, or social media.

Recommendations
The next step with this data would be to work with large language modelling techniques and train a model based on the perplexity of the texts, with the intention of increasing the applicability and accuracy of the deployed model.

Ethical Assessment
Some ethical considerations include the undisclosed method for procuring the essays featured in the data. At this point, assumptions are made that the essays were attained from public access sources or with consent. Also, it is very important to be able to tell the difference between human and AI-generated content because AI-generated content raises the risk of plagiarism and spreading misinformation, as well as copyright infringement and inconsistent quality. The primary ethical consideration of this project is the prevention of the misuse of AI-generated content, which could be perpetrated either through plagiarism in academic and professional settings or through the distribution of misinformation through channels of influence.

References

Gerami, S. (2024, January 10). *Ai vs human text*. Kaggle. https://www.kaggle.com/datasets/shanegerami/ai-vs-human-text/data

Hetler, A. (2023, September 21). *Pros and cons of ai-generated content: TechTarget*. WhatIs? https://www.techtarget.com/whatis/feature/Pros-and-cons-of-AI-generated-content

Matani, D. (2023, September 27). *Challenges of detecting AI-generated text*. Medium. https://towardsdatascience.com/challenges-of-detecting-ai-generated-text-6d85bf779448

White, M. (2023, July 8). *A brief history of generative AI*. Medium. https://matthewdwhite.medium.com/a-brief-history-of-generative-ai-cb1837e67106

Appendix 1

Evaluation Metrics of Sentence Length and Variability Models

Dummy test accuracy: 0.626

	Train Accuracy	Test Accuracy
Logistic regression	0.5984	0.5964
Naïve Bayes	0.5648	0.5639
Random Forest	0.7211	0.7022
Support Vector Classifier	0.6697	0.6662

Appendix 2

Evaluation Metrics of TF-IDF Models on Abbreviated Test/Train Data

Dummy test accuracy: 0.626

	Train Accuracy	Test Accuracy	Precision	Recall
Random Forest	1.0	0.9892	0.9978	0.9735
Support Vector	0.9986	0.9785	0.9819	0.9631
Naïve Bayes	0.9452	0.9372	0.9894	0.8423