Measuring the Impact of Explainability-Based Token Replacement on LLM-Generated Text

Hadi Mohammadi[1 [0000−0003−0860−9200]], Anastasia Giachanou[1 [0000−0002−7601−8667]], and Ayoub Bagheri[1 [0000−0001−6366−2173]]

Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands
{h.mohammadi, a.giachanou, a.bagheri}@uu.nl

Abstract. Generative AI models, especially Large Language Models (LLMs), have made great progress in producing human-like text. However, they sometimes exhibit patterns that make AI-generated content detectable compared to human writing. This study aims to explore how the explainability of models can help make AI-generated text less detectable. Specifically, we use Explainable AI techniques, particularly SHAP, to identify influential tokens that distinguish AI-generated text from human-written content. We then apply four token replacement methods: (i) replacing tokens with the most similar words used by humans, (ii) replacing tokens with similar words considering part-of-speech tagging, (iii) using generative models like GPT-4o mini to generate replacements; and (iv) using generative models with additional information about the genre of the text to make replacements. By implementing these modifications, we evaluate how changes in the tokens affect the explainable features and the overall detectability of the text. Our results show that certain token replacement strategies can reduce the detectability of AI-generated text, as evidenced by decreases in classification accuracy and changes in evaluation metrics. This research provides insights into the relationship between explainable features and text generation, with potential applications in content creation, AI model refinement, and natural language processing.

Keywords: Explainable AI · SHAP · Token Replacement · AI-Generated Text · Natural Language Processing

1 Introduction

Generative AI models, especially Large Language Models (LLMs) like GPT-x, have revolutionized text production by generating content that is very similar to human writing [2]. However, AI-generated text can still exhibit detectable patterns that differentiate it from human-authored material [11]. For instance, in the CLIN33 Shared Task dataset [12], AI-generated texts often adopt a clear, structured tone focused on objective reporting, as in "*The Mars rover Spirit has uncovered new clues about the existence of water on the red planet...*" In contrast, human-authored content might use expressive language to add layers

of meaning, such as "*The Hubble Space Telescope has a new, resonant date with destiny. NASA has set Sept. 11, 2008, as the day...*" The human text introduces subtle storytelling, which makes it feel more engaging. Similarly, when it comes to advocacy-related content, AI-generated text typically relies on strong, formulaic calls to action and standardized hashtags, as in "*Urging everyone to stand up against #SexualHarassment in the workplace! #ZeroTolerance for such behavior.*" Meanwhile, human-written content often shows a more personal, reflective tone, as seen in "*PSA as apparently some people are unsure: no one comes forward about sexual harassment for attention.*" This contrast shows how human-authored texts tend to bring individual perspective, while AI-generated texts, though effective in clarity and structure, can feel formulaic.

This difference between AI and human-generated text is especially important in cases where AI content needs to feel natural and undetectable to be effective and trustworthy. For example, when AI creates summaries for scientific papers or news articles, it's crucial that these summaries read as if they were written by human experts to keep them credible and easy to read [7]. Similarly, in tools that assist with creative writing, the AI suggestions should blend smoothly with the author's unique style [5]. In settings like hospitals, where AI may help generate discharge summaries, it's essential for these texts to feel clear and human-like to ensure accurate and understanding communication with patients [10]

While much of the current literature focuses on detecting AI content to prevent misinformation and defend content authenticity, our study looks at the equally important task of making AI-generated text sound more natural. There are other ways to label AI content, such as digital watermarking or metadata tagging [6], but these options might not work well or even be desirable in all cases. This highlights the need for AI to produce text that is naturally more human-like, especially in contexts where authenticity is essential. This study explores how we can help reduce the detectability of AI-generated text by using Explainable AI (XAI) techniques, specifically SHapley Additive exPlanations (SHAP) [8]. We identify key tokens that make AI-generated text distinguishable and analyze the effects of replacing these influential tokens. In particular, we aim to address research questions, first about whether explainable models can help make AI-generated text less detectable and, second, to what extent explainable features change when the AI-generated text changes.

2 Data

In this study, we used data from the CLIN33 Shared Task on the Detection of Text Generated by Large Language Models [4]. The dataset contains both AI-generated text and human-written text across multiple genres, such as news , tweets, and reviews, in both Dutch and English languages. In particular, the dataset consists of:

– **Human-Generated Text**: Texts collected from authentic sources, including newspapers, social media, and literary works.

– **AI-Generated Text**: Texts generated using LLMs like ChatGPT and GPT-4, based on prompts designed to mimic the styles of the human-generated texts.

Tables 1 provide an overview of the data distribution for the Dutch and English detection tasks, respectively.

Table 1: Data distribution for English and Dutch detection tasks [4]

English Dataset			
	Human-generated	AI-generated	Total
News	200	200	400
Twitter	200	200	400
Reviews	200	200	400
Dutch Dataset			
	Human-generated	AI-generated	Total
News	200	200	400
Twitter	200	200	400
Reviews	200	200	400

We did an 80-20 split on the combined data, resulting in 1,920 samples for training and 480 samples for testing. The split was randomized, and due to the balanced nature of the original dataset, the distribution of labels (AI-generated vs. human-written) remained balanced in both the training and test sets. Also we did the following preprocessing steps on the text data:

– **Lowercasing**: Converted all text to lowercase for uniformity.
– **Removal of URLs**: Eliminated any URLs present in the text using regular expressions.
– **Removal of Special Characters**: Removed punctuation and non-alphabetic characters to focus on textual content.
– **Tokenization**: Split text into individual tokens (words) using NLTK's word tokenizer [**?**].
– **Stop Word Removal**: Removed common stop words appropriate for each language using NLTK's stopword lists for English and Dutch.
– **Reconstruction**: Reconstructed the processed tokens back into text for vectorization.

3 Methodology

We propose a methodology that uses explainability techniques to enhance the undetectability of AI-generated text. Our approach involves training a classifier

to distinguish between AI-generated and human-written texts, interpreting the classifier's predictions to identify influential tokens, and then applying token replacement strategies to make the AI-generated texts less detectable. The code and results are publicly available in the GitHub repository[1].

We vectorized the preprocessed text data using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization, extracting the top 1,000 features to represent each text as a numerical vector. TF-IDF helps in highlighting important words in the text by scaling down the impact of commonly used words and scaling up the impact of less frequent, but potentially more informative, words.

We trained an XGBoost classifier [3], an efficient implementation of gradient-boosted decision trees, to distinguish between AI-generated and human-written texts. The classifier was trained using the default parameters with the evaluation metric set to logistic loss. The model's performance was evaluated using standard classification metrics, including precision, recall, F1-score, and accuracy.

We used SHAP [8] to integrate the explainability component into our methodology. SHAP values help understand the contribution of each feature (in this case, tokens) to the model's predictions. We used the TreeExplainer provided by SHAP, which is specifically designed for tree-based models like XGBoost.

The SHAP value for each token represents its average marginal contribution across all possible subsets of the tokens. By analyzing these values, we identified the tokens that have the most significant impact on the model's decision to classify a text as AI-generated. We calculated the mean absolute SHAP value for each token across all samples in the test set, ranked the tokens based on these values, and identified the most influential tokens for further analysis.

Based on the explainability analysis, we developed the following four token replacement strategies to modify the AI-generated texts and reduce their detectability. Each strategy focuses on replacing the most influential tokens identified by SHAP values.

1. **Strategy A: Human Similar Word Replacement (HSR)**
   In this strategy, we replaced the most influential tokens with the most similar words used by humans. We trained a Word2Vec model [9] on the human-written texts to learn word embeddings that capture semantic relationships between words. For each token to be replaced, we calculated the cosine similarity between its embedding and the embeddings of other words in the vocabulary.
2. **Strategy B: Part-of-Speech Replacement (PSR)**
   This strategy is an extension of Strategy A, adding grammatical consistency by considering part-of-speech (POS) tags. We performed POS tagging using NLTK's POS tagger [1]. For each token, we identified its POS tag and found similar words using the Word2Vec model as before.
3. **Strategy C: GPT-4 Replacement (GPTR)**
   In this strategy, we used a generative model, GPT-4o mini, to suggest replacements for the identified tokens. For each token, we constructed a prompt

[1] `https://github.com/hadimh93/Enhancing-AI-Generated-Text-Undetectability`

that asked the model to replace the token with a more human-like word in the given text.
   - Prompt: `f"Replace the token '{token}' with a more human-like word in the following text: '{text}'"`
4. **Strategy D: Genre-Specific GPT-4 Replacement (GGPTR)**
   Strategy D enhanced Strategy C by providing additional context about the genre of the text to GPT-4o mini. The prompt included genre-specific information to guide the model in generating more appropriate replacements consistent with the style and vocabulary typical of the genre.
   - Prompt: `f"Replace the token '{token}' with a more human-like word in the following {genre} text: '{text}'"`

4 Results

Before applying any token replacement strategies, we trained the classifier on the original AI-generated and human-written texts. Figure 1 shows the performance of the model across different languages and genres.

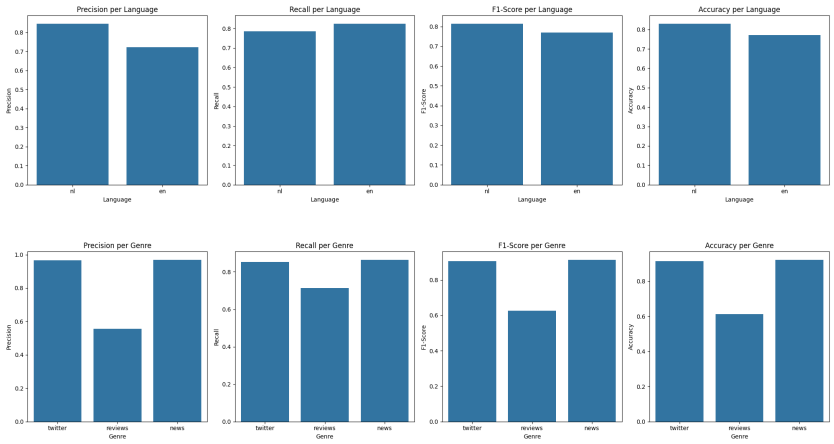

Fig. 1: Performance of the model across different languages (top) and genres (bottom).

Using SHAP values, we identified the tokens that most significantly contribute to the model's ability to detect AI-generated text. Figure 2 displays the top 20 tokens based on their mean absolute SHAP values. The x-axis represents the average magnitude of each token's contribution to the model's predictions—a higher value indicates a greater influence on the classification decision.

The most influential token was *mr"*, with a mean absolute SHAP value of 3.16, significantly higher than the others. While this might seem unexpected, it

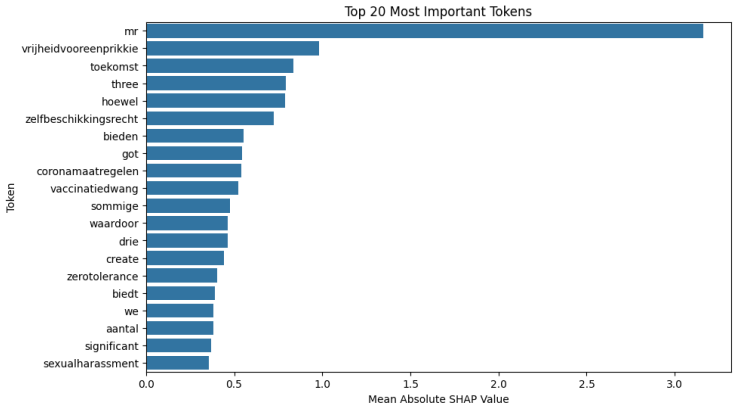

Fig. 2: Top 20 most influential tokens based on SHAP values

suggests that the model has learned to associate the usage of *mr"* differently in AI-generated versus human-written texts. This could be due to AI models using formal titles in contexts where human writers might not, leading to detectable patterns. Other highly influential tokens include Dutch words like *vrijheidvooreenprikkie"* (freedom for a jab), *toekomst"* (future), and *"hoewel"* (although). These domain-specific terms relate to topics such as vaccination and personal freedoms, which are prevalent in recent discourse. Their prominence indicates that the model captures content-related differences between AI-generated and human-written texts. The presence of both English and Dutch tokens among the top influencers reflects the bilingual nature of the dataset and shows that the model effectively captures language-specific features contributing to text classification. A detailed list of the most significant tokens and their mean absolute SHAP values is provided in the GitHub Repository.

4.1 Impact of Token Replacement Strategies

We applied the four token replacement strategies to the AI-generated texts in the test set and evaluated the impact on the model's performance. The results are summarized in Table 2.

As observed, each token replacement strategy resulted in a decrease in the model's accuracy compared to the original texts. Strategy HSR reduced the accuracy from 0.80 to 0.73, showing that the modified AI-generated texts became less detectable. Strategy PSR achieved an accuracy of 0.76, while Strategies GPTR and GGPTR both reached 0.78 accuracy. The decrease in the classifier's accuracy shows that the strategies were effective in making the AI-generated texts less detectable. By replacing the most influential tokens identified by SHAP values with more human-like words, the modified texts better mimic human writing patterns, thereby reducing the model's ability to correctly classify them as AI-generated.

Table 2: Classification Performance on AI-generated class After Applying Token Replacement Strategies

Strategy	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Original texts	0.78	0.81	0.79	0.80
Strategy HSR	0.74	0.65	0.69	0.73
Strategy PSR	0.75	0.73	0.74	0.76
Strategy GPTR	0.77	0.78	0.77	0.78
Strategy GGPTR	0.77	0.76	0.76	0.78

This reduction in detectability occurs because the classifier relies heavily on specific tokens to distinguish between AI-generated and human-written texts. When these tokens are altered or replaced, the features that the model uses for prediction are disrupted.The results indicate that Strategy HSR was more effective than the other strategies in making AI-generated texts less detectable, suggesting that simple replacement of influential tokens with similar words used by humans—without considering grammatical consistency or context—introduces variations that disrupt the model's ability to classify the texts correctly.

5 Discussion

Our study shows that explainability techniques can be effectively used to identify features that contribute to the detectability of AI-generated texts. By applying targeted token replacement strategies based on SHAP analysis, we can modify AI-generated texts to make them less detectable by classifiers. Our study shows that using explainability techniques, specifically SHAP values, can effectively identify the tokens that contribute most to the detectability of AI-generated text. By targeting these influential tokens for replacement, we can change the text in a way that reduces the model's ability to detect AI-generated content.

The token replacement strategies showed varying degrees of effectiveness. Strategy **HSR** resulted in the most decrease in classification accuracy, showing an improvement in the undetectability of the AI-generated texts. This suggests that simple replacement of influential tokens with similar words used by humans—without considering grammatical consistency or context—introduces variations is more effective.

Strategy **PSR** showed a moderate improvement over the original accuracy but was less effective than Strategy HSR. This indicates that maintaining grammatical consistency may preserve some of the patterns the classifier relies on, making the modified texts more detectable than those altered by Strategy HSR. Strategies **GPTR** and **GGPTR** (GPT-4 Replacement Strategies) achieved results similar to Strategy PSR but did not outperform it. While these strategies aim to produce more contextually appropriate and coherent texts, they may inadvertently retain stylistic features that the classifier uses for detection, resulting in a smaller decrease in accuracy compared to Strategy HSR. The findings sug-

gest that introducing unexpected variations through simple token replacement is more effective in evading detection by the classifier than making contextually or grammatically consistent changes.

6 Conclusion and Future Work

In this study, we explored how explainability techniques can be used to make AI-generated text less detectable. By employing SHAP values to identify the most influential tokens contributing to the classifier's decisions, we developed targeted token replacement strategies to modify AI-generated texts. Our findings showed that replacing these tokens with similar words used by humans without considering grammatical consistency (Strategy HSR) was the most effective approach in reducing the model's ability to detect AI-generated content. The results also show the dynamic relationship between explainable features and text generation.

For future work, we plan to incorporate automatic metrics such as BLEU, and ROUGE to evaluate the naturalness and readability of the modified AI-generated texts. This will help determine whether the strategies not only make the texts less detectable but also maintain or enhance their quality from a human perspective. Exploring advanced explainability techniques and alternative models could further improve our understanding of the features contributing to detectability and inform more effective strategies for text modification.

Furthermore, we intend to develop real-time applications that implement these token replacement strategies in live text generation scenarios. This could involve integrating the methodology into AI writing assistants or chatbots to produce content that is both high-quality and less susceptible to detection by classifiers. Testing the robustness of our strategies across different classification models will also be a focus, ensuring that the approach remains effective in diverse settings. By continuing to refine these techniques, we hope to enhance the naturalness and undetectability of AI-generated content, contributing valuable insights to the fields of artificial intelligence and computational linguistics.

Acknowledgments

We would like to thank the organizers of the CLIN33 Shared Task for providing the dataset used in this study.

References

1. Bird, S., Klein, E., Loper, E.: Categorizing and tagging words. Natural Language Processing with Python **179** (2009)
2. Brown, T.B., Mann, B., Ryder, N., et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems **33**, 1877–1901 (2020)
3. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)

4. Fivez, P., Daelemans, W., Van de Cruys, T., et al.: The clin33 shared task on the detection of text generated by large language models. Computational Linguistics in the Netherlands Journal **13**, 1–15 (2023)
5. Gero, K.I., Long, T., Chilton, L.B.: Social dynamics of ai support in creative writing. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–15 (2023)
6. Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., Goldstein, T.: A watermark for large language models. arXiv preprint arXiv:2301.10226 (2024)
7. Krishna, K., Song, Y., Karpinska, M., Wieting, J., Iyyer, M.: Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. Advances in Neural Information Processing Systems **36** (2024)
8. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems **30**, 4765–4774 (2017)
9. Mikolov, T.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
10. Patel, S.B., Lam, K.: Chatgpt: the future of discharge summaries? The Lancet Digital Health **5**(3), e107–e108 (2023)
11. Sadasivan, V.S., Kumar, A., Balasubramanian, S., Wang, W., Feizi, S.: Can ai-generated text be reliably detected? arXiv preprint arXiv:2303.11156 (2023)
12. Weller, O., Bisazza, A., van Noord, G.: The clin33 shared task on classifying machine-generated text. Computational Linguistics in the Netherlands Journal **13**, 201–221 (2023)