Literature Survey

Multilingual Content Integrity and Authorship Intelligence Platform:
An Ensemble Transformer Approach for AI-Generated Text Detection,
Semantic Plagiarism Identification, and Adaptive Content Humanization

1. Introduction

In the recent past, large language models have proven to be quite good at generating texts that look like they have been written by humans. With tools like ChatGPT, Gemini, and Claude, it is possible to generate essays, reports, and even research papers within a matter of seconds. This has brought about a number of genuine concerns, especially within the educational, journalistic, and publishing communities, since it is no longer easy to differentiate between texts written by humans and those written by machines. Plagiarism, too, has taken a different form, with many people paraphrasing texts from existing sources, a factor that has made it difficult for the common plagiarism detection tools to be useful.

The problem we aim to solve is a combination of three different but related issues, which include detecting whether a piece of text has been generated by a machine, detecting whether a piece of text has been plagiarized, and finally, rewriting the text to make it look like it is written by a human. Each of these areas has witnessed a considerable amount of research within the recent two to three years, and this paper is a discussion of the relevant papers within this field. The discussion will be divided into four different areas, which include benchmarks and data for AI text detection, methods for detecting texts generated by machines, sentence-level semantic similarity for detecting plagiarized texts, and finally, methods for rewriting texts to make them look like they have been written by humans.

2. Benchmarks and Datasets for AI Text Detection

A recurring issue with this research area has been the lack of standardised evaluation.

Existing detectors have typically been evaluated on small, in-house test sets, which

makes it difficult to compare results. This issue has been addressed by Dugan et al. with their shared test suite RAID, which includes more than six million documents generated by eleven different language models, eight different subject domains, and eleven different adversarial attack strategies [1]. What makes RAID stand out, though, is the sheer diversity of the test suite. It includes documents from eleven different language models, eight different subject domains, and eleven adversarial attack strategies. This means detectors have to generalise, rather than overfit to a particular model or writing style. The test suite was launched at ACL 2024, and since then, it has become the standard for evaluating new detectors.

Whilst RAID is primarily focused on the English language, the M4 dataset introduced by Wang et al. covers the multilingual dimension of this task.

M4 covers multiple generators, multiple languages, since AI-generated content is not limited to the English language. Their results show how poorly detectors perform when they have been exclusively trained on English language data, which led to our research.

decision to include XLM-RoBERTa in the detection ensemble.

Another popular dataset is the HC3 – Human ChatGPT Comparison Corpus, created by Guo et al. [3]. This dataset contains human-generated answers and ChatGPT responses for various question-answering domains. The paired nature of the data makes it convenient to train classifiers. However, it only contains data from a single generator and hence is best used as a supplement to other datasets like RAID and M4.

Recently, the FAIDSet framework provided a fine-grained labelling scheme, which moves beyond the traditional human vs. AI paradigm [11]. In the FAIDSet

framework, the traditional two-class labelling scheme is replaced by a three-class scheme. In the three-class scheme, a document can have parts of it written by a human and other parts written by a machine. This is more realistic, as AI tools are often used to write certain parts of a document while the rest is written by the human. The multi-task learning approach along with contrastive objectives also showed promising results for this three-class labelling scheme.

3. AI-Generated Text Detection Methods

The dominant strategy for detecting AI-generated text is to fine-tune a transformer model on labeled examples. He et al. proposed a model called DeBERTa, which incorporates a disentangled attention mechanism to separately model content and position information. This design gives DeBERTa an edge over other models for tasks requiring fine-grained token-level understanding. It has consistently performed at the top for natural language understanding tasks. For our project, DeBERTa-v3-large is used as the principal classifier for the detection ensemble. Tang et al. gave a comprehensive overview of detection strategies, which include statistical methods, neural classifiers, and watermark-based detection. The study revealed that no single strategy is superior to others. Statistical methods, such as perplexity scores, are effective for older generators, but they are less effective for newer models, which generate more coherent text. Neural classifiers, on the other hand, have higher accuracy but are less robust to test distributions that differ from the training distribution. This further corroborates our strategy of using an ensemble of four different architectures of transformers, rather than relying on a specific one.

The study by Sadasivan et al. posed critical questions on the inherent reliability of AI-based text detectors [4]. Their experimental study proved that paraphrasing attacks could lead to a significant reduction in detection accuracy, almost reaching a level of randomness. They proposed that, as the quality of language models is enhanced, the statistical differences between human-written and machine-generated texts will diminish, making it a challenging task for detection. This study served as a wake-up call for the field, which directly impacts our choice to include adversarial robustness tests.

Regarding the multilingual approach, Hlavnova and Pikuliak's study examined fine-tuning for machine-generated text detection across languages for the SemEval-2024 Task 8 [10]. Their study revealed that fine-tuning a LoRA version of the RoBERTa model, combined with a majority voting approach for language-specific heads, could lead to competitive performance without requiring the training of models for individual languages. Their study reinforced the effectiveness of parameter-efficient fine-tuning for multilingual detection, which is consistent with our approach using XLM-RoBERTa-large models trained on the M4 corpus.

4. Semantic Similarity for Plagiarism Detection

The traditional plagiarism detection techniques involve using n-gram overlap or fingerprinting. These techniques work well for exact copy plagiarism but do not work for paraphrased content. Foltynek, Meuschke, and Gipp reviewed the various plagiarism detection techniques in their comprehensive paper [12]. They have categorized plagiarism detection techniques into string-based, syntax-based, semantic-based, and citation-based techniques. They found that although the semantic-based approach is promising, it is also the least developed plagiarism detection technique. This project extends the work in the context of plagiarism detection by using sentence embeddings as the basis for the plagiarism detection module.

The basis for our work on plagiarism detection using embeddings is Sentence-BERT, which was introduced by Reimers and Gurevych in their paper [6]. They extended the traditional BERT approach to obtain fixed-size embeddings for sentences. These embeddings were achieved by using Siamese and triplet networks. These embeddings can be compared using the cosine similarity function. This approach is much faster in comparison to the traditional approach, in which the time complexity is quadratic in the number of documents to be compared. Gao, Yao, and Chen extended the work on sentence embeddings by introducing SimCSE, which is a contrastive learning approach to obtain better aligned and more uniform embeddings [7]. The unsupervised version of SimCSE uses dropout as a form of minimal data augmentation. This approach is simple and effective. In the supervised version, natural language inference is used to push the entailment examples closer and the contradiction examples further apart. We use SimCSE and Sentence-BERT embeddings to improve the recall of our plagiarism detection approach.

5. Text Rewriting and Content Humanization

The concept of rewriting text to evade detection is not novel, but Krishna et al. brought this concept into sharp focus in their work on DIPPER, a discourse-level paraphraser [8]. DIPPER is a paragraph-level paraphraser with two parameters: lexical diversity and order diversity. These parameters control the degree of change in the text's vocabulary and sentence order, respectively. Their work found that even moderate levels of paraphrasing were effective in evading most detectors but also found retrieval-based approaches to be effective in defending against paraphrasing. For our project, DIPPER is one of several rewriting models in our humanization module. The retrieval-based approach also influenced our choice to combine detection and semantic similarity search.

Another concept related to authorship style transfer is the ability to modify writing style while preserving content. Research on authorship style transfer using policy optimization techniques [13] found reinforcement learning to be effective in guiding a language model to mimic an author's writing style. Though our project does not involve authorship style transfer, the iterative feedback loop in our humanization module follows the same principle as reinforcement learning: the AI detection score is our reward function guiding our rewriting process to generate more human-like output.

Our approach to humanization involves three sequence-to-sequence models in a fallback chain: Flan-T5-XL as the first attempt, PEGASUS-large as an alternative with its abstractive summarization capabilities, and Mistral-7B fine-tuned with QLoRA as the most capable but also most resource-intensive option. The iterative feedback loop checks the output against the detection ensemble. If the AI score is still above the threshold, our system increases the diversity parameters and attempts rewriting again. This feedback loop is what differentiates our approach from one-shot paraphrasing tools.

6. Summary of Findings

There are a few themes that can be derived from the review. The first is that the space of AI-based text detection is advancing rapidly, and the availability of large-scale benchmark datasets such as RAID and M4 allows for more thorough evaluation of

approaches than was ever possible two years ago. The second is that no single approach to detection is sufficient on its own and that ensemble methods, using a variety of architectures and training methods, offer the best chance of generalization. The third is that semantic similarity detection methods have advanced to a point where paraphrase-level plagiarism can be reliably detected, although the computational complexity of such methods is a problem for very large datasets. The fourth is that text rewriting methods have become sophisticated enough to avoid detection by most methods, which is a problem and an opportunity for the humanization use case.

Our system incorporates all four of these strands. The detection module uses an ensemble of four transformer-based classifiers, informed by the benchmarking work of Dugan et al. and the multilingual results of Wang et al. The plagiarism module uses a combination of fast methods and semantic embedding techniques such as those of Reimers and Gurevych and Gao et al. The humanization module uses the text rewriting techniques of Krishna et al. and adds a feedback mechanism to treat detection as a quality gate.

References

[[1] Dugan, L., Hwang, A., Trhlik, F., et al., "RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors," Proceedings of the Association for Computational Linguistics (ACL), 2024. arXiv:2405.07940.

[2] Wang, Y., et al., "M4: Multi-generator, Multi-domain, Multi-lingual Black-Box Machine-Generated Text Detection," 2024. arXiv:2305.14902.

[3] Guo, B., et al., "HC3: Human ChatGPT Comparison Corpus," Hello-SimpleAI, Hugging Face, 2023.

[4] Sadasivan, V.S., et al., "On the Reliability of AI-Text Detectors," 2023. arXiv:2304.02819.

[5] Tang, R., et al., "Detecting Machine-Generated Text: A Critical Survey," 2023. arXiv:2303.07205.

[6] Reimers, N. and Gurevych, I., "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Proceedings of EMNLP, 2019. arXiv:1908.10084.

[7] Gao, T., Yao, X., and Chen, D., "SimCSE: Simple Contrastive Learning of Sentence Embeddings," Proceedings of EMNLP, 2021. arXiv:2104.08821.

[8] Krishna, K., et al., "Paraphrasing Evades Detectors of AI-Generated Text, but Retrieval is an Effective Defense," (DIPPER), 2023. arXiv:2303.13408.

[9] He, P., Liu, X., Gao, J., and Chen, W., "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," Proceedings of ICLR, 2021. arXiv:2006.03654.

[10] Hlavnova, E. and Pikuliak, M., “Fine-tuned LLMs for Multilingual Machine-Generated Text Detection,” SemEval-2024 Task 8. arXiv:2402.13671.

[11] “Fine-grained AI-generated Text Detection using Multi-task Auxiliary and Multi-level Contrastive Learning,” (FAIDSet), 2025. arXiv:2505.14271.

[12] Foltynek, T., Meuschke, N., and Gipp, B., “A Survey on Plagiarism Detection,” ACM Computing Surveys, 2019. arXiv:1703.05546.

[13] “Authorship Style Transfer with Policy Optimization,” 2024. arXiv:2403.08043.