Detecting AI-Generated Text:
A Survey

Abadila Alaktif
http://orcid.org/0009-0008-1010-706X
University Hassan II, Morocco

Meriyem Chergui
C3S, ENSEM, Morocco

Abdelkarim Ammoumou
C3S, ENSEM, Morocco

Gmira Faiq
C3S, ENSEM, Morocco

ABSTRACT

A thorough analysis of AI-generated text detection techniques is presented in this article. The authors offer a comprehensive examination of current detection techniques under three main headings: neural-based techniques that use deep learning models, statistics-based techniques that examine linguistic patterns, and watermarking-based techniques that embed identification markers. They tested how well these techniques worked with both general text and specialized academic material. The analysis identifies domain-specific problems in AI text detection by comparing the detection methods' performance across general text corpora and specialized scientific content. The results add to the growing corpus of research on AI content authentication, highlight existing constraints, and suggest future avenues for creating more effective detection tools.

INTRODUCTION

A significant challenge posed by large language models (LLMs) to the realm of text generation is the growing difficulty in distinguishing AI-generated output from that of a human. With the growing sophistication of these models, telling their output from that of a human is becoming a lot more difficult, leading to serious issues in digital media, journalism, and academia. As a result, the very existence

DOI: 10.4018/407609

of this trend necessitates stringent detection mechanisms to preserve the integrity of the content and safeguard intellectual property.

With evolution in the AI-geared text detection field-from watermarking methods introduced into the generation process to advanced neural network detection systems- watermarking has. AI language model developments happen so fast they outdate current detection methods and push novel theoretical frameworks and their practical uses.

Comparison and Our Contributions

Through various notable contributions inside the larger context of current methods, our work pushes the domain of artificial intelligence text detection forward.

1. **Comprehensive Analysis Framework:** We present a systematic examination of current detection methods in three major categories:
   - Watermarking-based approaches, including data-driven, model-driven, and post-processing techniques
   - Statistics-based methods, encompassing both white-box and black-box strategies
   - Neural-based detection systems, focusing on feature-based and pre-trained classifiers
2. **Comparative Assessment:** Our analysis provides critical insights into:
   - Performance variations between general and academic text detection
   - Strengths and limitations of different detection paradigms
   - Domain-specific challenges in scientific content authentication

Structure of the Article

This article is organized to provide a comprehensive overview of the field.

3. **Methodology (Section 2):** Details our systematic review approach following PRISMA guidelines:
   a. Identification and screening procedures for relevant literature
   b. Eligibility criteria and assessment methodology
   c. Data extraction and synthesis techniques
   d. Quality assessment framework for evaluating studies
4. **State of the Art (Section 3):** Presents a detailed analysis of existing detection methods in three primary categories:
   a. Watermarking-Based Methods: Examining embedded markers in text generation
   b. Statistics-Based Methods: Analyzing linguistic patterns and statistical signatures
   c. Neural-Based Methods: Exploring advanced deep learning approaches
5. **Results (Section 4):** Provides a thorough analysis of existing methods' effectiveness, including:
   a. Comprehensive discussion of current approaches
   b. Key findings from comparative studies
   c. Analysis of challenges and limitations
6. **Conclusion (Section 5**): Synthesizes findings across the detection landscape:
   a. Performance comparison of watermarking, statistical, and neural approaches
   b. Domain-specific detection challenges in scientific versus general content