

ABSTRACT

IDENTIFICATION OF MACHINE-GENERATED TEXT USING STYLOMETRIC AND CONTENT-BASED FEATURES

by

Putu Annand Kurnia Iswari
20/457776/PA/19814

The rapid advancement of artificial intelligence (AI), in particular large language models (LLM), has brought about a fundamental shift in the manner in which humans interpret and produce content that is based on text. Large Language Models (LLMs), such as the Generative Pre-trained Transformer 3 (GPT-3), have the ability to generate text that is hauntingly similar to the writing that is produced by humans. As a result, the study to use artificial intelligence for text classification and analysis, especially LLM-generated writing, has conducted in the past few year. It resulted an advanced approaches using a very complex language models that capable to distinguish the two texts. However, as a deep learning method works in black box, it is hard to identify the pattern to distinguish the two text. In this research, the identification of text characteristicstics will take stylometric and content-based features. Moreover, this research also build a machine learning model from these features to classify the human-generated text and machine-generated text, using support vector machine, logistic regression, and random forest. From the experiment, stylometric features alone demonstrate remarkable effectiveness, particularly when paired with the Random Forest classifier, achieving the highest metrics with an accuracy of 0.9181. The combination of stylometric and content-based features generally enhances classifier performance, though not significantly improving beyond its results with stylometric features alone. The high performance achieved with stylometric features indicates that writing style characteristics are sufficient for effective classification.

Keyword: Machine-Generated Text Detection, Stylometric Analysis, Content-Based Features, Natural Language Processing