

INTISARI

ANALISIS TINGKAT AKURASI METODE PENDETEKSIAN PLAGIARISME IDE DENGAN MENGGUNAKAN *YAKE* DAN *SENTENCE TRANSFORMER*

Oleh:

Salsabila Laily Rahma

19/441237/SV/16589

Penggunaan *Artificial Intelligence* dengan teknik *unsupervised learning* dapat berguna dalam pendeteksian plagiarisme ide karena dapat mengidentifikasi kemiripan dan perbedaan antara dokumen teks tanpa memerlukan data berlabel atau pelatihan khusus. Plagiarisme ide melibatkan penyisipan ringkasan dari satu dokumen teks ke dalam dokumen teks lainnya, sehingga membuatnya sulit terdeteksi menggunakan metode pendeteksian plagiarisme standar. Metode yang dikembangkan dalam penelitian ini bertujuan untuk mengatasi permasalahan masalah dalam deteksi plagiarisme ide. Penelitian ini mengembangkan metode untuk deteksi plagiarisme ide dan menguji tingkat akurasi level dokumen dari metode yang dikembangkan. Metode yang dikembangkan ini menggunakan pendekatan baru dengan memanfaatkan *library Python* yang mengimplementasikan *AI* pada teknik *unsupervised learning* yaitu metode *Yake* sebagai algoritma pengekstrak kata kunci dan *Sentence Transformer* sebagai algoritma untuk menghitung kemiripan teks pada *dataset PAN. Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN)* adalah kumpulan data yang bersifat *public* dan secara khusus dikembangkan untuk penelitian dalam bidang pendeteksian plagiarisme. *Dataset PAN* yang digunakan dalam penelitian ini adalah *dataset PAN13-14 summary obfuscation* dengan sebuah *ground truth* yang menjadi acuan dalam pengukuran akurasi dari metode yang dikembangkan. Hasil penelitian menunjukkan bahwa metode *Sentence Transformer* dengan *Yake* pada *threshold* 0.1 memiliki akurasi tertinggi untuk kategori *Plagiarized* dengan nilai *F-score* pada *dataset testing* dan *dataset training* secara berturut-turut adalah 0.3175 dan 0.3217, sementara metode *Sentence Transformer* dengan *threshold* 0.6 memiliki akurasi tertinggi untuk kategori *Non Plagiarized* dengan nilai *F-score* pada *dataset testing* dan *dataset training* secara berturut-turut adalah 0.8905 dan 0.8907.

Kata kunci: Kecerdasan Buatan, Deteksi Plagiarisme Ide, *Yake*, *Sentence Transformer*, *Python*, Akurasi, *Dataset PAN13-14*

ABSTRACT

ANALYSIS OF ACCURACY LEVEL IN IDEA PLAGIARISM DETECTION METHOD USING YAKE AND SENTENCE TRANSFORMER

By:

Salsabila Laily Rahma

19/441237/SV/16589

The utilization of Artificial Intelligence (AI) with unsupervised learning techniques can be beneficial in detecting idea plagiarism as it can automatically identify similarities and differences between textual documents without requiring labeled data or specialized training. Idea plagiarism involves inserting a summary from one text document into another, making it challenging to detect using conventional plagiarism detection methods. The method developed in this research aims to address the issues related to detecting idea plagiarism. This study develops a method for detecting idea plagiarism and evaluates the accuracy level of the developed method. The approach utilizes a novel approach by leveraging Python libraries implementing AI techniques, specifically Yake as a keyword extraction algorithm and Sentence Transformer as a text similarity computation algorithm on the PAN dataset (Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection). The PAN dataset is publicly available and specifically designed for research in the field of plagiarism detection. The dataset used in this study is the PAN13-14 summary obfuscation dataset, and it contains ground truth data used as a reference for measuring the accuracy of the developed method. The research findings indicate that the Sentence Transformer method with Yake at a threshold of 0.1 achieves the highest accuracy for the Plagiarized category, with F-score values of 0.3175 and 0.3217 on the testing and training datasets, respectively. On the other hand, the Sentence Transformer method with a threshold of 0.6 achieves the highest accuracy for the Non Plagiarized category, with F-score values of 0.8905 and 0.8907 on the testing and training datasets, respectively.

Keywords: Artificial Intelligence, Idea Plagiarism Detection, Yake, Sentence Transformer, Python, Accuracy, PAN13-14 Dataset