Analisis Tingkat Akurasi Metode Pendeteksian Plagiarisme Ide dengan Menggunakan *Yake* dan *Sentence Transformer*

Salsabila Laily Rahma[1], Umar Taufiq[1,*]
[1]Departemen Teknik Elektro dan Informatika, Sekolah Vokasi, Universitas Gadjah Mada;
salsabilalaily@mail.ugm.ac.id
[*]Korespondensi: umartaufiq8284@ugm.ac.id;

Abstract – 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. This research develops a method to address issues in 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

Intisari – 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. Penelitian ini mengembangkan metode untuk mengatasi permasalahan dalam 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*

I. PENDAHULUAN

Seiring dengan bertambahnya jumlah penggunaan teknologi internet dan ketersediaan data yang semakin meningkat, hal ini membuat banyak akademisi, peneliti hingga pelajar yang menggunakan data dari internet untuk kebutuhan pribadi. Dalam rangka pencegahan tindakan plagiarisme yang marak terjadi, seringkali para pengguna mengubah teks, mengganti istilah hingga menggunakan sinonim sebagai pengganti kata aslinya, melakukan parafrase, mengganti kalimat aktif menjadi kalimat pasif atau sebaliknya hingga menggunakan teknik lainnya. Hal ini menjadikan tindakan plagiarisme ide atau gagasan menjadi sebuah masalah yang sulit dideteksi menggunakan metode pendeteksian plagiarisme standar [1].

Plagiarisme ide merupakan tindakan mengutip ide orang lain sebagai miliknya sendiri tanpa menyebutkan sumbernya. Plagiarisme dalam dokumen teks yang bersifat tersembunyi (*obfuscated*) tidak dapat terdeteksi oleh metode deteksi plagiarisme yang hanya menyandarkan kinerjanya pada deteksi kesamaan teks saja. Inti permasalahan yang akan diselesaikan dalam penelitian ini adalah kesulitan dalam mendeteksi plagiarisme ide, terutama dalam dokumen teks yang memiliki *obfuscated plagiarism* yang sulit dideteksi oleh metode deteksi plagiarisme konvensional. Oleh karena itu, pengembangan metode deteksi plagiarisme ide dengan memanfaatkan teknologi informasi yang canggih perlu untuk dilakukan untuk mencegah plagiarisme dan menjaga keunikan sumber informasi. [2]

Pendeteksian plagiarisme mengacu pada serangkaian kegiatan di bidang teknologi informasi yang terdiri dari proses *preprocessing* dan normalisasi teks, representasi dokumen, hingga perhitungan kemiripan. Proses *preprocessing* dan normalisasi teks akan menghilangkan tanda baca, kata penghubung, *comments*, *blanks* dan *extra lines*. Proses

representasi dokumen adalah proses untuk merepresentasikan dokumen ke dalam bentuk yang dapat dibaca oleh sistem, dalam hal ini dapat dilakukan dengan perbandingan karakter *string* antara dua buah dokumen. Proses perhitungan kemiripan dapat dilakukan dengan mengimplementasikan algoritma untuk mendeteksi kasus plagiarisme dengan didasarkan pada *matrix* perangkat lunak [3].

Di dalam sistem pendeteksian plagiarisme, terdapat metodologi yang diterapkan yaitu *Artificial Intelligence* (AI) yang memanfaatkan teknik *unsupervised learning.* Hal ini dapat berguna dalam pendeteksian plagiarisme karena dapat menyimulasikan pemikiran manusia menggunakan data sehingga dapat menghasilkan perilaku tanpa campur tangan manusia dalam menanggapi prediksi tersebut. Keuntungan teknik *unsupervised learning* berupa dapat menggunakan data yang tidak berlabel untuk mencari struktur, pola, *cluster*, pengelompokan atau observasi terkait dalam data *input* [4].

Penelitian ini menghadirkan kontribusi baru dalam upaya pencegahan plagiarisme dengan mengusulkan penggunaan kombinasi algoritma yang menggunakan pendekatan teknik pembelajaran tanpa pengawasan (*unsupervised learning*) sebagai metode pendeteksian plagiarisme ide yang akan diuji pada *dataset summary obfuscation*. Pendekatan ini dianggap baru karena belum secara luas dikembangkan dan diterapkan dalam deteksi plagiarisme ide. Hal inilah yang mendorong penulis untuk melakukan penelitian dengan harapan akan memberikan kontribusi penting dan solusi yang lebih efektif dalam mengembangkan metode pendeteksian plagiarisme ide.

Yake merupakan sebuah *library Python* yang menggunakan pendekatan teknik pembelajaran tanpa pengawasan (*unsupervised learning*) sehingga tidak perlu dilatih pada kumpulan dokumen tertentu. *Yake* dapat digunakan untuk mengotomatiskan proses ekstraksi kata kunci secara cepat, dimana dapat mengidentifikasi kata kunci penting dan paling relevan dengan akurasi tinggi dalam teks. *Yake* akan menghasilkan daftar kata kunci *final* yang telah diurutkan berdasarkan skor relevansinya [5].

Sentence Transformer merupakan sebuah *framework Python* yang menggunakan pendekatan teknik pembelajaran tanpa pengawasan (*unsupervised learning*) sehingga tidak membutuhkan data *training* berlabel. *Sentence Transformer* dapat digunakan untuk penyematan kalimat dan teks dengan lebih dari 100 bahasa. Penyematan ini dapat dibandingkan dengan *cosine similarity* untuk menemukan kalimat dengan arti yang mirip. Hal ini dapat berguna untuk mencari kemiripan kalimat dalam teks [6].

Berdasarkan metode pendeteksian plagiarisme yang telah tersedia secara *public*, maka terdapat sebuah dasar pemikiran pada penelitian ini untuk mengembangkan metode pendeteksian plagiarisme ide dengan memanfaatkan *Yake* dan *Sentence Transformer*, serta menguji tingkat akurasi level dokumen pada metode pendeteksian plagiarisme ide yang dikembangkan pada dataset *PAN13-14 summary obfuscation.* Dataset *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 *PAN13-14* yang digunakan terdiri dari dokumen teks berbahasa Inggris yang terbagi menjadi *dataset testing set* dan *dataset training set*. Setiap *dataset* ini memiliki pasangan dokumen teks yang terdiri dari dokumen asli atau sumber (*source*) dan dokumen yang dicurigai melakukan plagiasi terhadap dokumen asli (*suspicious*), serta dokumen *ground truth* yang berisi daftar pasangan dokumen *source* dan *suspicious* beserta status *plagiat* atau *bukan plagiat* [7].

Dalam penelitian ini, dibutuhkan pengujian metode pendeteksian plagiarisme dengan penetapan nilai *threshold* dari 0.1 - 0.9 untuk mengetahui tingkat akurasi yang dihasilkan. Penelitian ini juga menggunakan *ground truth* yang berguna untuk memvalidasi *threshold* yang mendekati kebenaran. *Yake* akan dikembangkan untuk mendeteksi plagiarisme dengan menghitung persamaan kata kunci dalam suatu pasangan dokumen teks. Metode *Sentence Transformer* juga dapat digunakan untuk mendeteksi kemiripan teks secara mandiri maupun dalam kombinasi dengan *Yake*. Perhitungan plagiarisme pada metode *Sentence Transformer* secara mandiri dilakukan dengan cara memanfaatkan model *transformer* yang dapat menghasilkan representasi vektor dari penyematan teks yang diambil dari pasangan dokumen sumber (*source*) dan dokumen yang mencurigakan (*suspicious*).

Representasi vektor yang dihasilkan dari *Sentence Transformer* akan dibandingkan menggunakan *cosine similarity* untuk menentukan tingkat kemiripan antara kalimat-kalimat. Sementara perhitungan plagiarisme pada metode *Sentence Transformer* dalam kombinasi dengan *Yake* dilakukan dengan cara mengotomatiskan proses ekstraksi kata kunci antar pasangan dataset, lalu kata kunci tanpa skor akan dimasukkan ke dalam model *transformer*. Hasilnya adalah representasi vektor yang dapat dibandingkan menggunakan *cosine similarity* untuk mengidentifikasi kalimat-kalimat dengan arti yang mirip sebagai indikasi adanya plagiarisme. 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.

II. DASAR TEORI

Adanya metode pendeteksian plagiarisme *Yake* dan *Sentence Transformer* serta tersedianya *dataset PAN 13-14 summary obfuscation* yang telah dilengkapi dengan *ground truth*, memberikan dasar pemikiran untuk mengembangkan metode pendeteksian plagiarisme ide serta menguji tingkat akurasi level dokumen pada metode yang dikembangkan melalui pemanfaatan teknik *artificial intelligence.* Metode pendeteksian plagiarisme ini merupakan bentuk implementasi dari *artificial intelligence*, khususnya pada teknik *unsupervised learning. Yake* tidak perlu dilatih pada kumpulan dokumen tertentu untuk keperluan ekstraksi kata kunci pada teks, sementara *Sentence Transformer* tidak

membutuhkan data *training* berlabel untuk keperluan mencari kemiripan kalimat pada teks [2, 4, 5, 6, 8]. *Dataset PAN 13-14 summary obfuscation* dapat menjadi data yang digunakan pada penelitian untuk mengetahui tingkat akurasi level dokumen karena terdiri dari dataset *testing* dan *training* yang mana setiap *dataset* memiliki dokumen *source* dan *suspicious*. [7].

Penelitian terkait penerapan metode ekstraksi kata kunci dan perhitungan kemiripan semantik dengan penyematan dan perbandingan kalimat telah banyak dilakukan untuk optimisasi guna menguji akurasi dalam kinerjanya. Penelitian [9], sebuah sistem dibuat untuk melakukan ekstraksi frasa kunci dari data *social media* menggunakan *Sentence Transformer* dengan model *deep learning BERT* sebagai model pembelajaran mesin yang menggunakan proses penyematan *tweet* dan kemudian menggunakan matrik kemiripan untuk menghitung skor kemiripan. Selain itu, dilakukan penggabungan semua frasa kunci dalam satu dokumen dengan penerapan nilai ambang batas untuk proses pemeringkatan sehingga didapatkan frasa kunci terbaik. Penelitian [10], sebuah sistem dibuat untuk mengembangkan metode untuk *semantic textual similarity* (STS) dengan cara *SBERT* untuk mendapatkan penyisipan kalimat yang bermakna secara semantik yang dapat dibandingkan menggunakan *cosine similarity*.

Penelitian [11], sebuah sistem dibuat untuk implementasi relevansi teks guna menemukan dokumen topik tertentu melalui serangkaian proses yang terdiri dari *text preprocessing*, proses ekstraksi kata kunci dengan algoritma *Porter Stemming*, pembobotan kata kunci dengan cara perhitungan relevansi teks dengan *cosine similarity* yang akan menghasilkan nilai antara 0 dan 1, dimana nilai yang semakin mendekati 1 maka akan semakin tinggi kemiripan kedua dokumen. Penelitian [12], sebuah sistem bernama FacTeR-Check dibuat untuk melakukan pengecekan fakta atau mengekstraksi informasi mendalam tentang penyebaran hoaks dengan cara pengambilan *tweet* yang terkait hoaks untuk memfilter *tweet* yang paling relevan dengan *XLM-RoBERTa Transformer* untuk mengevaluasi kemiripan semantik dan ekstraksi kata kunci dengan *KeyBERT* serta menemukan kemiripan dengan menggunakan *Named-Entity Recognition*. Penelitian [13], sebuah sistem dibuat untuk mendeteksi plagiarisme pada dokumen teks dengan metode algoritma *Jaccard Similarity* pada dokumen *source* dan dokumen *suspicious* pada *dataset PAN* tahun 2013 dan 2014.

Penelitian-penelitian di atas hanya berfokus pada metode ekstraksi kata kunci dan perhitungan kemiripan semantik dengan penyematan dan perbandingan kalimat. Sejauh ini belum ada penelitian yang berfokus pada gabungan dari metode ekstraksi kata kunci dengan perhitungan kemiripan dengan *transformer* kalimat untuk mengukur akurasi level dokumen dalam hal deteksi plagiarisme ide pada *dataset* yang diakui dan digunakan oleh peneliti secara umum. Penelitian ini menggunakan pendekatan baru pada pengembangan metode pendeteksian plagiarisme ide yaitu dengan menggabungkan konsep ekstraksi kata kunci dan perhitungan kemiripan teks. Metode ini unik dan belum pernah diteliti sebelumnya sehingga patut dikembangkan dalam penelitian sehingga menjadikan penelitian ini berpotensi memberikan hasil yang inovatif untuk keperluan analisis dan pengembangan metode di masa yang akan datang.

Dengan adanya teknologi serta ketersediaan *dataset* yang dapat mendukung penelitian, maka hal ini menciptakan sebuah ide atau dasar pemikiran pada penelitian ini yang bertujuan untuk untuk mengembangkan metode pendeteksian plagiarisme ide dengan memanfaatkan *Yake* dan *Sentence Transformer*, serta menguji tingkat akurasi level dokumen pada metode pendeteksian plagiarisme ide yang dikembangkan pada dataset *PAN13-14 summary obfuscation.* Pengukuran akurasi level dokumen menggunakan formula *Recall* (*Rd*), *Precision* (*Pd*), dan *F-score* (*Fd*) [14].

Penelitian ini menggunakan metode yang menjadi acuan karena sudah dikembangkan oleh peneliti sebelumnya yaitu metode *Sentence Transformer*, sementara metode yang dimodifikasi dalam penelitian adalah metode *Yake Similarity*, *Yake Similarity* (Modif.#1), dan gabungan metode *Sentence Transformer* dengan tambahan *Yake*.

Metode *Yake Similarity* adalah sebuah metode yang dikembangkan dari algoritma *yake* dalam rangka menghitung kemiripan pada teks. Secara *default, yake* merupakan sebuah algoritma pengekstrak kata kunci. Penelitian ini mengusung pendekatan yang unik yaitu dengan menghitung persamaan kata kunci yang dihasilkan oleh *yake* pada dua dokumen teks *source* dan *suspicious*. Nilai kemiripan didapatkan dari hasil perhitungan jumlah kata kunci yang sama antara 2 dokumen (*irisan*) / jumlah kata kunci keseluruhan pada 2 dokumen. Berdasarkan data hasil pengujian setelah metode ini diimplementasikan, masih ditemukan adanya kelemahan pada metode ini yaitu nilai akurasi maksimal hanya berada pada 0.5 sehingga kelemahan metode ini perlu ditingkatkan.

Metode *Yake Similarity* (*Modif*.#1) adalah sebuah metode penyempurnaan dari metode *Yake Similarity* guna meningkatkan nilai akurasinya. Penelitian ini menghitung persamaan kata kunci yang dihasilkan oleh *yake* pada dua dokumen teks *source* dan *suspicious*. Nilai kemiripan didapatkan dari hasil perhitungan jumlah kata kunci yang sama antara 2 dokumen (*irisan*) / gabungan kata kunci antara 2 dokumen (*union*). Perhitungan kemiripan pada metode ini berhasil meningkatkan akurasi dari metode *Yake* sebelumnya, dengan rata-rata kenaikan akurasi pada dataset *testing* sebesar 0.001721797107 dan pada dataset *training* sebesar 0.001528264308.

Metode *Sentence Transformer* adalah sebuah metode yang menjadi sebuah acuan atau referensi karena metode ini sudah dikembangkan oleh peneliti sebelumnya. Metode ini mampu menemukan kemiripan kalimat dalam teks dengan teknik *sentence embeddings* dan perbandingan dengan *cosine similarity.* Dalam implementasinya, metode ini akan memasukkan kalimat atau paragraf dari dokumen *source* dan *suspicious* ke dalam model *transformer* sehingga menghasilkan sebuah *source embeddings* dan *suspicious embeddings* berupa vektor yang berukuran tetap. Kemudian kedua vektor ini dibandingkan dengan *cosine similarity. Cosine similarity* merupakan metode bawaan (*built-in*) dari *Sentence Transformer Library.*

Persamaan (1) dan (2) menunjukkan formula yang diterapkan dalam metode *cosine similarity* untuk membandingkan tingkat kemiripan dua buah vektor yang merupakan representasi dari teks. Persamaan tersebut menjadi landasan utama dalam metodologi penelitian ini.

$$Cosine\ Similarity(A,B) = (A.B)/(||A|| * ||B||) \quad (1)$$

$$\frac{(a_1b_1+a_2b_2+a_3b_3+\ldots+a_{384}b_{384})}{\sqrt{a_1{}^2+a_2{}^2+a_3{}^2+\ldots+a_{384}{}^2}\sqrt{b_1{}^2+b_2{}^2+b_3{}^2+\ldots+b_{384}{}^2}} \quad (2)$$

Metode *Sentence Transformer* dan *Yake* adalah sebuah metode yang dikembangkan dalam penelitian ini. Metode ini merupakan contoh implementasi dari gabungan konsep ekstraksi kata kunci dan perhitungan kemiripan teks sebagai sebuah pendekatan baru yang belum ada sebelumnya. Dalam implementasinya, metode ini akan melakukan proses ekstraksi kata kunci dengan *yake* lalu sistem akan mengambil kata kunci tanpa skor dari dokumen *source* dan *suspicious* untuk dimasukkan ke dalam model *transformer* sehingga menghasilkan *source embeddings* dan *suspicious embeddings* berupa vektor yang berukuran tetap untuk dibandingkan dengan *cosine similarity*. Dengan demikian, pengembangan metode ini diharapkan mampu menghasilkan nilai akurasi lebih baik karena *sentence transformer* dapat langsung memproses kata kunci paling relevan yang telah dihasilkan oleh *Yake*.

Penelitian ini dimaksudkan untuk menguji tingkat akurasi level dokumen pada metode pendeteksian plagiarisme ide yang dikembangkan. Hal ini cukup sulit dilakukan karena terdapat proses pengelompokan (*clustering*) *dataset* sesuai dengan *ground truth* yang harus diselesaikan sebelum proses perhitungan nilai kemiripan pada metode pendeteksian plagiarisme ide dilakukan. Penggunaan *dataset PAN 13-14 summary obfuscation* yang sudah dilengkapi dengan *ground truth* memberikan kelebihan yaitu dapat digunakan untuk memvalidasi *threshold* yang mendekati kebenaran.

Tabel 1 menunjukkan statistik *dataset PAN13-14* yang digunakan dalam penelitian. Hal ini bertujuan agar sistem dapat menguji setiap pasangan *dataset* (*source document* dan *suspicious document*) seperti yang telah ditentukan pada *ground truth* agar tidak menimbulkan banyak kombinasi jika membandingkan dokumen *source* dan dokumen *suspicious* secara keseluruhan.

Tabel 1. Statistik dataset pada *PAN13-14 summary obfuscation* menurut Potthast dkk [15]

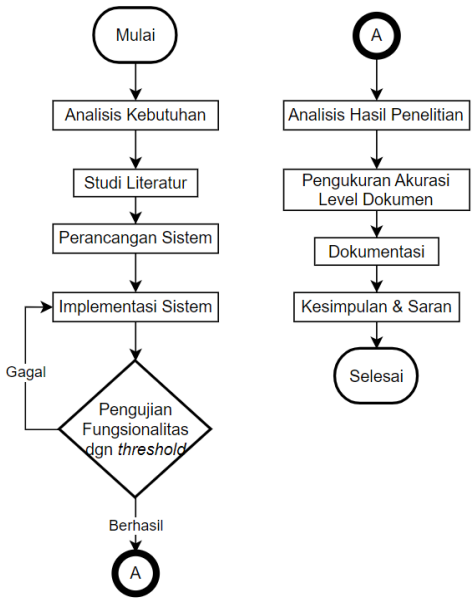
Documents	***Training set***		***Testing set***	
	suspicious	***source***	***suspicious***	***source***
Non-Plagiarized	947	-	949	-
Plagiarized	238	-	236	-
Total	1185	237	1185	237

III. METODOLOGI

Metode penelitian yang dilakukan berfokus pada pengembangan metode dan pengukuran akurasi level dokumen pada metode yang dikembangkan dengan memanfaatkan *Yake* dan *Sentence Transformer*.

A. Analisis Kebutuhan yaitu dengan melakukan riset mengenai kebutuhan dan spesifikasi sistem, serta kebutuhan perangkat lunak dan perangkat keras yang digunakan untuk mendukung penelitian.
B. Tahap perancangan yaitu dengan membuat rencana dan rancangan mengenai program sistem yang terdiri atas dua kegiatan yaitu studi literatur dan perancangan sistem. Studi literatur dilakukan dengan mengumpulkan literatur-literatur yang berkaitan dengan perangkat lunak. Perancangan sistem dilakukan dengan mempersiapkan dataset, instalasi dan konfigurasi sistem, perancangan struktur direktori dan desain sistem yang akan digunakan pada penelitian.
C. Implementasi teknik ekstraksi kata kunci dan perhitungan kemiripan teks dengan Yake dan Sentence Transformer. Langkah-langkah yang perlu dijalankan pada tahapan ini adalah seperti pembuatan file berisi kata kunci, pengelompokan file berisi kata kunci dan dataset PAN, pembuatan file berisi nilai kemiripan dengan metode pendeteksian plagiarisme, serta pengujian *threshold* pada metode pendeteksian plagiarisme.
D. Pengujian sistem dilakukan dengan cara pengujian fungsionalitas sistem untuk memastikan bahwa proses menerima *input* dan menghasilkan *output* sudah berfungsi sebagaimana mestinya dan pengujian terkait dengan pengukuran tingkat akurasi level dokumen pada metode pendeteksian plagiarisme dengan menghitung *recall, precision* dan *F-score* berdasarkan data jumlah pasangan dokumen yang diidentifikasi oleh sistem sebagai *plagiat* maupun *non plagiat* (baik yang sukses maupun gagal terdeteksi).

Gambar 1 menunjukkan langkah-langkah proses penelitian dengan digambarkan menggunakan diagram alir yang memvisualisasikan secara sistematik urutan proses yang dilakukan dalam penelitian. Proses yang tertera pada penelitian ini dilakukan sesuai dengan metodologi yang telah ditetapkan. Metodologi yang mendasari penelitian ini adalah metode *waterfall.* Pendekatan ini dipilih sebagai kerangka kerja yang paling sesuai untuk mencapai tujuan penelitian dengan efektif dan efisien, memastikan kelancaran pelaksanaan serta konsistensi hasil yang diperoleh dalam penelitian.



Gambar 1. Skema rangkaian yang digunakan pada penelitian

IV. HASIL DAN PEMBAHASAN

Untuk melakukan evaluasi terhadap keakuratan metode pendeteksian plagiarisme ide dengan menggunakan *Yake* dan *Sentence Transformer*, dilakukan pengujian dengan menetapkan nilai ambang batas *threshold* mulai dari 0.1 - 0.9 untuk menentukan apakah suatu dokumen dikatakan plagiat atau tidak. Pengujian ini dilakukan untuk dataset *PAN13-14 summary obfuscation* dengan jumlah dokumen *source* dan dokumen *suspicious* serta *ground truth* yang dapat digunakan. Pada dataset *testing*, terdiri dari jumlah dokumen *source* sebanyak 3169 dan jumlah dokumen *suspicious* sebanyak 1826. Pada dataset *training*, terdiri dari jumlah dokumen *source* sebanyak 3230 dan jumlah dokumen *suspicious* sebanyak 1827. Semua eksperimen dilakukan dengan menggunakan bahasa pemrograman *Python* dengan menguji 1185 pasangan dokumen *source* dan dokumen *suspicious* pada dataset *testing* dan dataset *training* berdasarkan data *ground truth*.

Selanjutnya perbandingan akurasi tingkat dokumen antara metode *Yake* dengan *Sentence Transformer* dapat dijelaskan sebagai berikut. Sesuai dengan ekspektasi seperti yang tampak pada Tabel 2 - 5, hasil penelitian yang diperoleh menunjukkan bahwa pada kategori *plagiarized*, tingkat akurasi tertinggi diperoleh pada kombinasi metode *Yake* dan *Sentence Transformer* (*threshold* 0.1), sebagaimana dibuktikan dengan adanya nilai *F-score* tertinggi pada *dataset testing* dan *dataset training* secara berturut-turut adalah 0.3175 dan 0.3217. Sementara pada kategori *non plagiarized*, tingkat akurasi tertinggi diperoleh pada metode *Sentence Transformer* (*threshold* 0.6) dengan nilai *F-score* pada *dataset testing* dan *dataset training* yaitu sebesar 0.8905 dan 0.8907.

Akurasi yang diperoleh dari metode *Sentence Transformer* dengan tambahan *Yake* dapat meningkatkan akurasi dalam deteksi plagiarisme ide. Hal ini dikarenakan *Yake* mampu memberikan kata kunci paling relevan dari teks dengan akurasi tinggi sehingga kinerja *Sentence Transformer* dalam mencari kemiripan dalam kalimat akan semakin baik. Hasilnya adalah berupa sistem dapat mengidentifikasi jumlah pasangan dokumen lebih banyak daripada metode lain yang dikembangkan dalam penelitian.

Hasil penelitian ini juga menunjukkan bahwa akurasi pada metode *Yake* tanpa adanya metode *Sentence Transformer* perlu ditingkatkan. Metode *Yake* menunjukkan ketidakseimbangan performa metode. Hal ini dapat dilihat pada akurasi yang terlalu rendah pada kategori *Plagiarized* dan akurasi yang terlalu tinggi pada kategori *Non Plagiarized.* Oleh karena itu, metode ini tidak dijadikan sebagai hasil kesimpulan meskipun memiliki akurasi tertinggi. Berbeda halnya dengan metode *Sentence Transformer* yang mampu membedakan atau mengklasifikasikan kasus *Plagiarized* dan *Non Plagiarized* dengan lebih baik berdasarkan variasi *threshold*. Sementara pada metode *Sentence Transformer* dengan tambahan metode Yake mampu meningkatkan akurasi level dokumen menjadi lebih baik.

Tabel 2 menunjukkan perbandingan *F-score* pada metode pendeteksian plagiarisme yang diujikan pada *dataset testing* pada kategori *Plagiarized.* Berdasarkan hasil pengujian, akurasi level dokumen tertinggi pada kategori *Plagiarized* yaitu terdapat pada threshold 0.1 dengan menggunakan metode kombinasi antara *Yake* dan *Sentence Transformer.* Hal ini dapat dilihat bahwa pada kasus *plagiarized,* nilai Fd tertinggi pada *dataset testing* adalah 0.3175.

Tabel 2. Akurasi pada dataset testing (Plagiarized)

Threshold	***Yake***	***Sentence***	***Sentence Yake***
0.1	0.2321	0.2967	0.3175
0.2	0.0574	0.2496	0.2712
0.3	-	0.1838	0.1269
0.4	-	0.1188	0.0752
0.5	-	0.0785	0.04
0.6	-	0.0331	0.0402
0.7	-	0.0084	0.0402
0.8	-	-	0.0324
0.9	-	-	0.0247

Tabel 3 menunjukkan perbandingan *F-score* pada metode pendeteksian plagiarisme yang diujikan pada *dataset training* pada kategori *Plagiarized.* Berdasarkan hasil pengujian, akurasi level dokumen tertinggi pada kategori *Plagiarized* yaitu terdapat pada threshold 0.1 dengan menggunakan metode kombinasi antara *Yake* dan *Sentence Transformer.* Hal ini dapat dilihat bahwa pada kasus *plagiarized,* nilai Fd tertinggi pada *dataset training* adalah 0.3217.

Tabel 3. Akurasi pada dataset training (Plagiarized)

Threshold	*Yake*	*Sentence*	*Sentence Yake*
0.1	0.2369	0.3194	0.3217
0.2	0.0084	0.2454	0.2655
0.3	-	0.1715	0.1798
0.4	-	0.1088	0.1194
0.5	-	0.0775	0.0923
0.6	-	0.0569	0.0781
0.7	-	0.0084	0.056
0.8	-	-	0.056
0.9	-	-	0.056

Tabel 4 menunjukkan perbandingan *F-score* pada metode pendeteksian plagiarisme ide yang diujikan pada *dataset testing* pada kategori *Non Plagiarized*. Berdasarkan hasil pengujian, metode *Sentence Transformer* dengan *threshold* 0.6 terbukti memiliki akurasi tertinggi dalam mendeteksi pasangan dokumen yang diidentifikasi sebagai *bukan plagiat* dengan nilai *F-score* yaitu 0.8905.

Tabel 4. Akurasi pada dataset testing (Non Plagiarized)

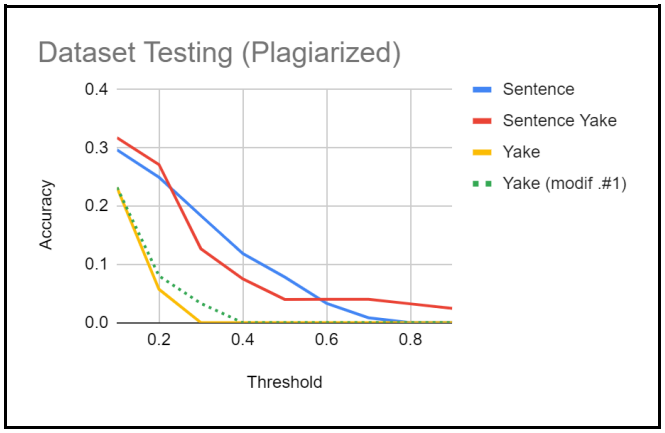
Threshold	*Yake*	*Sentence*	*Sentence Yake*
0.1	0.8917	0.5458	0.4552
0.2	0.8918	0.7673	0.7707
0.3	0.8894	0.849	0.8583
0.4	0.8894	0.879	0.8831
0.5	0.8894	0.8889	0.8868
0.6	0.8894	0.8905	0.8873
0.7	0.8894	0.8898	0.8873
0.8	0.8894	0.8894	0.888
0.9	0.8894	0.8894	0.8886

Tabel 5 menunjukkan perbandingan *F-score* pada metode pendeteksian plagiarisme ide yang diujikan pada *dataset training* pada kategori *Non Plagiarized.* Berdasarkan hasil pengujian, metode *Sentence Transformer* dengan *threshold* 0.6 terbukti memiliki akurasi tertinggi dalam mendeteksi pasangan dokumen yang diidentifikasi sebagai *bukan plagiat* dengan nilai *F-score* yaitu 0.8907.

Gambar 2 menunjukkan grafik perbandingan *F-score* pada metode pendeteksian plagiarisme ide yang diujikan pada *dataset testing* pada kategori *Plagiarized.* Dalam hal ini, metode yang dimodifikasi dalam penelitian adalah metode *Sentence Transformer* dan *Yake* (merah), *Yake* (kuning), dan *Yake modif.#1* (hijau), sementara metode yang menjadi acuan karena sudah dikembangkan oleh peneliti sebelumnya adalah metode *Sentence Transformer* (biru).

Tabel 5. Akurasi pada dataset *training* (*Non Plagiarized*)

Threshold	*Yake*	*Sentence*	*Sentence Yake*
0.1	0.8877	0.543	0.4539
0.2	0.8888	0.747	0.762
0.3	0.8884	0.8393	0.8602
0.4	0.8884	0.8738	0.8877
0.5	0.8884	0.8873	0.8882
0.6	0.8884	0.8907	0.8884
0.7	0.8884	0.8888	0.8887
0.8	0.8884	0.8884	0.8887
0.9	0.8884	0.8884	0.8887



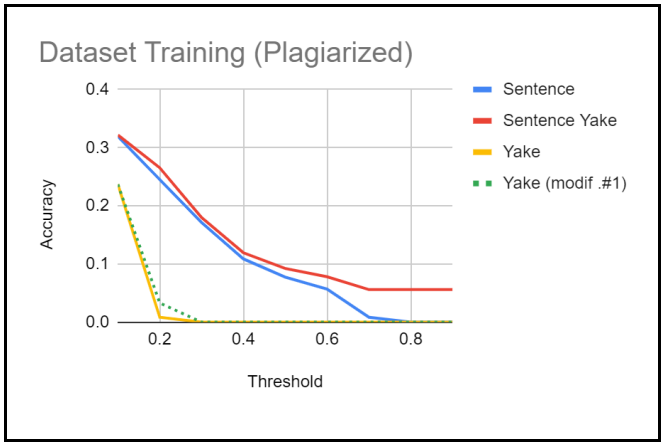
Gambar 2. Perbandingan Akurasi pada *Dataset Testing* untuk kategori *Plagiarized*

Gambar 3 menunjukkan grafik perbandingan *F-score* pada metode pendeteksian plagiarisme ide yang diujikan pada *dataset training* pada kategori *Plagiarized.* Dalam hal ini, metode yang dimodifikasi dalam penelitian adalah metode *Sentence Transformer* dan *Yake* (merah), *Yake* (kuning), dan *Yake modif.#1* (hijau), sementara metode yang menjadi acuan karena sudah dikembangkan oleh peneliti sebelumnya adalah metode *Sentence Transformer* (biru).

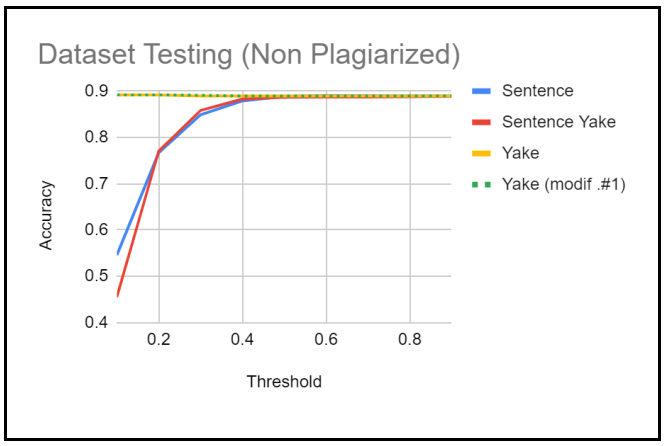
Gambar 4 menunjukkan grafik perbandingan *F-score* pada metode pendeteksian plagiarisme ide yang diujikan pada *dataset testing* pada kategori *Non Plagiarized.* Dalam hal ini, metode yang dimodifikasi dalam penelitian adalah metode *Sentence Transformer* dan *Yake* (merah), *Yake* (kuning), dan *Yake modif.#1* (hijau), sementara metode yang menjadi acuan karena sudah dikembangkan oleh peneliti sebelumnya adalah metode *Sentence Transformer* (biru).

Gambar 5 menunjukkan grafik perbandingan *F-score* pada metode pendeteksian plagiarisme ide yang diujikan pada *dataset training* pada kategori *Non Plagiarized.* Dalam hal ini, metode yang dimodifikasi dalam penelitian adalah metode

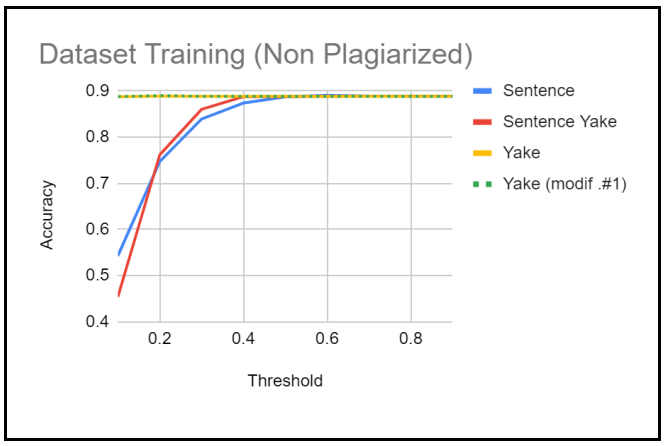
Sentence Transformer dan *Yake* (merah), *Yake* (kuning), dan *Yake modif.#1* (hijau), sementara metode yang menjadi acuan karena sudah dikembangkan oleh peneliti sebelumnya adalah metode *Sentence Transformer* (biru).



Gambar 3. Perbandingan Akurasi pada *Dataset Training* untuk kategori *Plagiarized*



Gambar 4. Perbandingan Akurasi pada Dataset *Testing* untuk kategori *Non Plagiarized*



Gambar 5. Perbandingan Akurasi pada Dataset *Training* untuk kategori *Non Plagiarized*

V. SIMPULAN

Pada penelitian ini telah dibuat sebuah rancang bangun sistem pendeteksian plagiarisme ide dengan metode *Yake* dan *Sentence Transformer* untuk mengukur akurasi level dokumen pada *dataset PAN13-14 summary obfuscation.* Berdasarkan hasil penelitian, dapat disimpulkan bahwa semakin rendah nilai batas *threshold* yang ditetapkan dalam suatu pengujian, maka jumlah pasangan dokumen yang terdeteksi sebagai *plagiat* dalam sistem akan semakin meningkat, sementara jumlah pasangan dokumen yang terdeteksi sebagai *bukan plagiat* akan semakin menurun.

Berdasarkan penelitian yang telah dilakukan, maka diperoleh beberapa saran dalam upaya pengembangan dan perbaikan penelitian di masa depan. Adapun saran yang dapat dilakukan untuk peneliti selanjutnya meliputi a.) melakukan penelitian dengan mengembangkan metode untuk meningkatkan akurasi level dokumen, b.) melakukan penelitian dengan menggunakan *dataset* selain berbahasa Inggris, c.) melakukan penelitian dengan menggunakan sampel *dataset* yang lebih besar.

REFERENSI

[1] V. Bhuyar and S.N. Deshmukh. 2023. “Analysis of Support Tools for Plagiarism Detection,” International Conference on Applications of Machine Intelligence and Data Analytics: ICAMIDA 2022, ACSR 105, pp. 38-46, 2023.

[2] Taufiq, Umar. 2023. Named Entity Recognition dan Dependency Parsing untuk Ekstraksi Konsep yang Lebih Baik dalam Deteksi Plagiarisme. Disertasi. Program Studi S3 Ilmu Komputer. Universitas Gadjah Mada. Yogyakarta.

[3] A. S. Bin-Habtoor and M.A. Zaher. 2012. “A Survey on Plagiarism Detection Systems,” International Journal of Computer Theory and Engineering, vol. 4, no. 2, pp. 1-5.

[4] H. Garg, J. M. Chatterjee, R. N. Thakur. 2023. “A Roadmap For Enabling Industry 4.0 By Artificial Intelligence,” USA: John Wiley & Sons, Inc. and Scrivener Publishing LLC.

[5] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, A. Jatowt. 2019. “Yake! Keyword Extraction from Single Documents using Multiple Local Features,” Information Sciences, vol. 509, 2020, pp. 257-289, Elsevier Inc.

[6] Reimers, Nils. 2022. “Sentence Transformer Documentation,” dilihat pada 6 Juni 2023 halaman situs https://www.sbert.net/

[7] M. N. Mansoora and M.S.H. Al-Tamimib, 2022. “ Computer-based plagiarism detection techniques: A comparative study,” 13th International Journal of Nonlinear Analysis and Applications (IJNAA), 2022, pp. 10-12.

[8] Patel, Ankur A., 2019. “Hands-On Unsupervised Learning Using Python: How to Build Applied Machine Learning Solutions from Unlabeled Data,” USA: O’Reilly Media, Inc. (Penerbit O’Reilly), pp. 3-26. dilihat pada 12 Juni 2023 halaman situs https://www.google.co.id/books/edition/Hands_On_Unsupervised_Learning_Using_Pyt/-SKJDwAAQBAJ?hl=en&gbpv=1&dq=unsupervised+learning+algorithm+AI&printsec=frontcover

[9] R. Devika, S. Vairavasundaram, C. S. J. Mahenthar, V. Varadarajan and K. Kotecha, "A Deep Learning Model Based on BERT and Sentence Transformer for Semantic Keyphrase Extraction on Big Social Data," in *IEEE Access*, vol. 9, pp. 165252-165261, 2021, doi: 10.1109/ACCESS.2021.3133651.

[10] Reimers, Nils and Gurenvych, Iryna. 2019. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” dilihat pada 6 Juni 2023 halaman situs https://arxiv.org/abs/1908.10084.

[11] D. Gunawan, C.A. Sembiring, M.A Budiman, "The Implementation of Cosine Similarity to Calculate Text Relevance Between Two Documents," 2018. IOP Conference Series: Journal of Physics (J.Phys.), pp. 1-6.

[12] A. Martin, J. Huertas-Tato, A. Huertas-García, G. Villar-Rodríguez, D. Camacho. 2022. "FacTeR-Check: Semi-automated fact-checking through semantic similarity and natural language inference," Knowledge Based Systems, vol. 251, pp. 1-17. Spain: Elsevier Inc.

[13] Agustian, Surya. 2021. "Pendekatan Semantik Dalam Deteksi Berbagai Tipe Plagiarisme Pada Dokumen Teks," Jurnal Teknik Informatika, vol.14, no.2, pp. 1-14, (file PDF). diunduh pada 13 Juni 2023 halaman situs https://journal.uinjkt.ac.id/index.php/ti/article/downloadSuppFile/22841/5618

[14] Potthast, M, Hagen, M, Gollub, T, Tippmann, M, Kiesel, J, Rosso, P, Stamatatos, E,dan Stein, B, 2013b, Overview of the 5th International Competition on Plagiarism Detection. In Forner, P, Navigli, R, Tufis, D, dan Ferro, N, editors, Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013, volume 1179 of CEUR Workshop Proceedings. CEUR-WS.org.

[15] Potthast, M, Gollub, T, Hagen, M, Tippmann, M, Kiesel, J, Rosso, P, Stamatatos, E, dan Stein, B, September 2013a. PAN13 Originality: Text Alignment. URL https://doi.org/10.5281/zenodo.3715980.