DETEKSI PLAGIARISME PADA NOVEL BERBAHASA INGGRIS MENGGUNAKAN *AUTHORSHIP ATTRIBUTION* BERBASIS *STYLOMETRY* DAN *SUPPORT VECTOR MACHINE* (SVM)

LAPORAN TUGAS AKHIR

Laporan ini Disusun untuk Memenuhi Salah Satu Syarat Memperoleh Gelar Sarjana Strata 1 (S1) pada Program Studi Teknik Infromatika Fakultas Teknologi Industri Universitas Islam Sultan Agung Semarang

DISUSUN OLEH:

MEY RINI RZ

32602100064

PROGRAM STUDI TEKNIK INFORTIKA

FAKULTAS TEKNOLOGI INDUSTRI

UNIVERSITAS ISLAM SULTAN AGUNG

2025

PLAGIARISM DETECTION IN ENGLISH NOVELS USING AUTHORSHIP ATTRIBUTION BASED ON STYLOMETRY AND SUPPORT VECTOR MACHINE (SVM)

FINAL PROJECT REPORT

Poposed to complete the requirement to obtain a bachelor's degree (S-1) at Informatics Engineering Departement of Industrial Technologgy Faculty Sultan Agung Islamic University

Arranged By:

MEY RINI RZ

32602100064

MAJORING OF INFORMATICS ENGINEERING

INDUSTRIAL TECHNOLOGY FACULTY

SULTAN AGUNG ISLAMIC UNIVERSITY

SEMARANG

2025

LEMBAR PENGESAHAN
TUGAS AKHIR

DETEKSI PLAGIARISME PADA NOVEL BERBAHASA INGGRIS MENGGUNAKAN *AUTHORSHIP ATTRIBUTION* BERBASIS *STYLOMETRY* DAN *SUPPORT VECTOR MACHINE* (SVM)

MEY RINI RZ
32602100064

Telah dipertahankan di depan tim penguji ujian sarjana tugas akhir
Program Studi Teknik Informatika
Universitas Islam Sultan Agung
Pada tanggal : 1 September 2025

TIM PENGUJI UJIAN SARJANA :

Imam Much Ibnu Subroto, ST, M.SC, Ph.D NIK. 210600017 (Ketua Penguji)		16-09-2025
Ir. Sri Mulyono, M.Eng NIK. 210616049 (Anggota Penguji)		19-09-2025
Badieah, ST, M.Kom NIK. 210615044 (Pembimbing)		01-10-2025

Semarang, ...

Mengetahui,

Kaprodi Teknik Informatika
Universitas Islam Sultan Agung

Moch Taufik, ST, MIT
NIK. 210604034

SURAT PERNYATAAN KEASLIAN TUGAS AKHIR

Yang bertanda tangan dibawah ini :

Nama : MEY RINI RZ

NIM : 32602100064

Judul Tugas Akhir : Deteksi Plagiarisme Pada Novel Berbahasa Inggris Menggunakan Authorship Attribution Berbasis Stylometry Dan Support Vector Machine (Svm)

Dengan bahwa ini saya menyatakan bahwa judul dan isi Tugas Akhir yang saya buat dalam rangka menyelesaikan Pendidikan Strata Satu (S1) Teknik Informatika tersebut adalah asli dan belum pernah diangkat, ditulis ataupun dipublikasikan oleh siapapun baik keseluruhan maupun sebagian, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka, dan apbila di kemudian hari ternyata terbukti bahwa judul Tugas Akhir tersebut pernah diangkat, ditulis ataupun dipublikasikan, maka saya bersedia dikenakan sanksi akademis. Demikian surat pernyataan ini saya buat dengan sadar dan penuh tanggung jawab.

Semarang, 2 oktober 2025

Yang Menyatakan,

Mey Rini Rz

PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH

Saya yang bertanda tangan dibawah ini :

Nama : MEY RINI RZ

NIM : 32602100064

Program Studi : Teknik Informatika

Fakultas : Teknologi industri

Dengan ini menyatakan Karya Ilmiah berupa Tugas akhir dengan Judul : Deteksi Plagiarisme pada Novel Berbahasa Inggris Menggunakan *Authorship Attribution* Berbasis *Stylometry* dan *Support Vector Machine* (Svm)

Menyetujui menjadi hak milik Universitas Islam Sultan Agung serta memberikan Hak bebas Royalti Non-Eksklusif untuk disimpan, dialihmediakan, dikelola dan pangkalan data dan dipublikasikan diinternet dan media lain untuk kepentingan akademis selama tetap menyantumkan nama penulis sebagai pemilik hak cipta. Pernyataan ini saya buat dengan sungguh-sungguh. Apabila dikemudian hari terbukti ada pelanggaran Hak Cipta/Plagiarisme dalam karya ilmiah ini, maka segala bentuk tuntutan hukum yang timbul akan saya tanggung secara pribadi tanpa melibatkan Universitas Islam Sultan agung.

Semarang, 2 Oktober 2025

Yang menyatakan,

KATA PENGANTAR

Puji syukur penulis panjatkan kepada Allah SWT, yang telah memberikan rahmat, taufik serta hidayah-Nya, sehingga Tugas Akhir dengan judul Deteksi Plagiarisme Pada Novel Berbahasa Inggris Menggunakan *Authorship Attribution* Berbasis *Stylometry* Dan *Support Vector Machine* (Svm) dapat terselesaikan.

Tanpa lupa penulis mengucapkan terima kasih untuk beberapa pihak yang telah membantu secara materi, pikiran, dan dukungan mental. Saya selaku penulis mengucapkan terima kasih kepada :

1. Rektor UNISSULA Bapak Prof. Dr. H. Gunarto, S.H., M.H., yang mengizinkan penulis menimba ilmu di kampus ini.
2. Dekan Fakultas Teknologi Industri Ibu Dr. Ir. Hj. Novi Marlyana, S.T., M.T., IPU., ASEAN.Eng.
3. Dosen Pembimbing Ibu Badie'ah, ST, M.Kom yang telah membimbing dan memberikan banyak nasehat serta saran.
4. Orang tua dan keluarga penulis yang telah membatu secara materi dan mengizinkan untuk menyelesaikan laporan ini.
5. Dan kepada semua pihak yang tidak dapat saya sebutkan satu persatu.

Penulis menyadari bahwa dalam penulisan laporan ini masih terdapat banyak kekurangan, untuk itu penulis mengharap kritik dan saran dari pembaca untuk sempurnanya laporan ini. Semoga dengan ditulisnya laporan ini dapat menjadi sumber ilmu bagi setiap pembaca.

Semarang,

Mey Rini Rz

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Plagiarisme pada novel berbahasa Inggris tidak hanya berupa penyalinan langsung, tetapi juga peniruan gaya penulisan (*paraphrase plagiarism*). Penelitian ini mengembangkan sistem deteksi berbasis *authorship attribution* dengan *stylometry*, *Support Vector Machine* (SVM), dan *Sentence*-BERT (SBERT). Data berupa 15 novel dari lima penulis klasik diproses melalui *preprocessing* dan chunking menjadi 1000, 5000, dan 10000 kata. Hasil pengujian menunjukkan akurasi SVM sebesar 84.38% (1000 kata), 82.50% (5000 kata), dan tertinggi 90.48% (10000 kata). Jane Austen konsisten mudah dikenali dengan f1-*score* 0.90, sementara Mary Shelley meningkat signifikan pada teks panjang (*recall* 1.00). Analisis SBERT menghasilkan skor kesamaan semantik 0.55–0.63, dengan nilai tertinggi juga pada Austen (0.63). Integrasi SVM dan SBERT terbukti saling melengkapi serta *stylometry* efektif mengenali gaya, sedangkan SBERT menangkap kesamaan makna. Dengan demikian, sistem mampu mendeteksi plagiarisme secara lebih akurat dan komprehensif.

Kata Kunci: Plagiarisme, *Stylometry*, *Authorship Attribution*, SVM, SBERT

ABSTRACT

Plagiarism in English novels is not limited to direct copying but also includes paraphrase plagiarism that imitates writing style. This study develops a detection system based on authorship attribution using stylometry, Support Vector Machine (SVM), and Sentence-BERT (SBERT). The dataset consists of 15 novels by five classic authors, processed through preprocessing and chunking into 1000, 5000, and 10000 words. Experimental results show SVM achieved accuracies of 84.38% (1000 words), 82.50% (5000 words), and the highest 90.48% (10000 words). Jane Austen was consistently well-identified with f1-scores 0.90, while Mary Shelley improved significantly on longer texts (recall 1.00). SBERT analysis produced semantic similarity scores ranging from 0.55 to 0.63, with the highest score also for Austen (0.63). The integration of SVM and SBERT proved complementary stylometry effectively captured writing style, while SBERT detected semantic meaning. Thus, the system enables more accurate and comprehensive plagiarism detection.

Keywords: Plagiarism, Stylometry, Authorship Attribution, SVM, SBERT

BAB I
PENDAHULUAN

1.1 Latar Belakang

Plagiarisme merupakan tindakan tidak etis yang melibatkan pengambilan karya orang lain tanpa memberikan atribusi yang semestinya. Dalam dunia akademik dan literatur, plagiarisme dapat merusak integritas penulis dan nilai orisinalitas suatu karya. Seiring dengan berkembangnya teknologi, penyebaran dan modifikasi teks melalui media digital menjadi lebih mudah, sehingga risiko plagiarisme pun meningkat, termasuk dalam karya sastra seperti novel berbahasa Inggris. Plagiarisme dalam karya sastra tidak selalu berbentuk salinan langsung, tetapi juga bisa berupa *paraphrase plagiarism* atau pencurian gaya menulis yang halus, yang lebih sulit dideteksi dengan metode konvensional(Maurya dkk., 2021).

Salah satu pendekatan yang berkembang untuk mendeteksi plagiarisme adalah dengan *authorship attribution* berbasis *stylometry*, yaitu metode yang menganalisis gaya linguistik dan kebiasaan penulisan individu. Gaya ini mencakup berbagai fitur seperti frekuensi kata, panjang kalimat, struktur gramatikal, dan pola tanda baca, yang secara statistik dapat merepresentasikan identitas penulis. *Authorship attribution* dapat digunakan untuk mengidentifikasi penulis suatu teks, bahkan jika teks tersebut telah dimodifikasi atau disamarkan (He dkk., 2024).

Lebih lanjut, penggabungan proses *preprocessing* seperti tokenisasi, ekstraksi fitur *stylometry*, dan penggunaan model klasifikasi seperti SVM telah dikembangkan dalam sistem deteksi plagiarisme modern. Penggunaan kombinasi *stylometry* dan SVM mampu mendeteksi plagiarisme bahkan pada teks hasil parafrase yang meniru gaya penulisan tertentu. Selain itu, analisis kemiripan menggunakan metode seperti cosine similarity dan model semantik modern (misalnya BERT) juga semakin melengkapi sistem ini dalam tahap validasi (El-Rashidy dkk., 2024).

Dalam penelitian ini, pendekatan *authorship attribution* akan digunakan untuk menganalisis kemiripan gaya penulisan dalam novel berbahasa Inggris, menggunakan fitur-fitur *stylometry* seperti panjang kalimat, frekuensi kata umum, hingga distribusi tanda baca. Model SVM akan dilatih pada kumpulan teks dari

beberapa penulis berbeda, kemudian diuji untuk mengidentifikasi potensi plagiarisme berdasarkan kesamaan gaya penulisan.

Dengan demikian, penerapan metode *stylometry* dan SVM dalam deteksi plagiarisme literatur fiksi membuka peluang baru dalam pengawasan keaslian karya tulis sastra. Sistem ini diharapkan dapat membantu akademisi, penerbit, maupun penulis untuk menjaga orisinalitas dan integritas karya mereka di era digital yang semakin kompleks.

1.2 Perumusan Masalah

1. Bagaimana mengidentifikasi penulis (*author*) dari potongan teks novel berbahasa Inggris menggunakan ciri-ciri gaya penulisan (*stylometry*) dan algoritma *Support Vector Machine* (SVM)?
2. Bagaimana mengembangkan sistem deteksi plagiarisme dengan memanfaatkan kemiripan gaya penulisan dan makna teks menggunakan fitur *stylometry*, *Sentence*-BERT?

Dengan rumusan masalah tersebut, penelitian ini dapat diharapkan memberikan pemahaman lebih mendalam terkait penerapan *machine learning* berbasis *stylometry* dan *semantic similarity* dalam mendeteksi kemungkinan adanya plagiarisme secara otomatis pada teks literatur panjang seperti novel.

1.3 Pembatasan Masalah

1. Penelitian ini hanya menggunakan novel berbahasa inggris sebagai objek analisis. Novel yang dianalisis diambil dari genre fiksi, dengan variasi panjang teks yang cukup besar.
2. Jumlah data yang digunakan dalam penelitian terdiri dari 5 penulis fiksi klasik berbahasa Inggris, masing-masing sebanyak 3 novel, sehingga total terdapat 15 novel sebagai data utama.
3. Plagiarisme yang dideteksi terbatas pada kemiripan gaya penulisan (*stylometry*) dan kemiripan semantik antar teks, bukan pencocokan isi secara literal atau *copy-paste* langsung.
4. Sistem tidak mendeteksi plagiarisme dalam skenario penerjemahan atau parafrase lintas bahasa.

1.4 Tujuan

Tujuan penelitian ini adalah untuk:

1. Mengidentifikasi penulis (*author*) dari potongan teks novel berbahasa Inggris dengan memanfaatkan ciri-ciri gaya penulisan (*stylometry*) dan menerapkan algoritma klasifikasi *Support Vector Machine* (SVM).
2. Mengembangkan sistem deteksi plagiarisme yang di ukur dari kemiripan antara teks yang akan di analisa dengan teks milik penulis yang telah teridentifikasi, melalui perbandingan gaya penulisan (*stylometry*) dan kemiripan makna *(semantic similarity*) menggunakan model SBERT.

Dengan tujuan tersebut, sistem yang dikembangkan dapat membantu dalam mengidentifikasi indikasi plagiarisme secara otomatis dalam teks panjang berbasis gaya penulisan penulis asli.

1.5 Manfaat

Manfaat dari penelitian ini adalah memberikan kontribusi pada pengembangan *authorship attribution* berbasis *stylometry* dan machine learning melalui penerapan algoritma *Support Vector Machine* (SVM), sekaligus menghadirkan sistem dasar untuk mendeteksi plagiarisme pada teks sastra dengan mengombinasikan analisis gaya penulisan dan kesamaan semantik menggunakan *Sentence*-BERT (SBERT). Penelitian ini diharapkan dapat menjadi referensi bagi pengembangan metode deteksi plagiarisme modern serta memperkuat upaya menjaga orisinalitas karya sastra di era digital.sastra.

1.6 Sistematika Penulisan

Penulisan laporan penelitian ini disusun dengan sistematika sebagai berikut:

BAB I PENDAHULUAN

Berisi latar belakang, rumusan masalah, pembatasan masalah, tujuan, manfaat, dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Berisi teori-teori dasar yang menjadi landasan penelitian, seperti *stylometry*, SVM, *Sentence*-BERT, serta beberapa penelitian terdahulu yang relevan.

BAB III METODE PENELITIAN

Menjelaskan tahapan penelitian mulai dari desain penelitian, pengumpulan data, *preprocessing*, ekstraksi fitur *stylometry*, pelatihan model klasifikasi, dan evaluasi deteksi plagiarisme.

BAB IV HASIL DAN PEMBAHASAN

Menyajikan implementasi program, hasil pelatihan model SVM, evaluasi performa model, serta hasil deteksi plagiarisme menggunakan *Sentence*-BERT.

BAB V KESIMPULAN DAN SARAN

Berisi kesimpulan yang didapatkan dari penelitian ini dan saran untuk penelitian selanjutnya yang dapat memperluas atau meningkatkan efektivitas metode yang digunakan.

BAB II
TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Penelitian dalam bidang *authorship attribution* berfokus pada proses identifikasi penulis suatu teks berdasarkan karakteristik khas gaya penulisan yang mereka miliki. Pendekatan ini menerapkan fitur-fitur linguistik yang disebut *stylometry*, seperti frekuensi penggunaan kata, panjang kalimat, tanda baca, hingga kompleksitas sintaksis. Fitur-fitur ini kemudian dianalisis menggunakan algoritma pembelajaran mesin (*machine learning*) untuk melakukan klasifikasi penulis.

Support Vector Machine (SVM) merupakan salah satu metode *machine learning* berbasis pembelajaran terawasi (*supervised learning*) yang banyak digunakan untuk klasifikasi maupun regresi. Algoritma ini diperkenalkan pertama kali oleh Vapnik pada tahun 1995 dan sejak itu menjadi salah satu teknik yang populer karena kemampuannya menghasilkan akurasi tinggi pada berbagai bidang aplikasi, mulai dari pengenalan pola, pengolahan citra, hingga analisis data teks. Prinsip dasar dari SVM adalah mencari sebuah *hyperplane* optimal yang berfungsi sebagai pemisah antar kelas dengan jarak maksimum terhadap titik data terdekat dari masing-masing kelas, yang dikenal sebagai *support vectors*. Dengan margin pemisah yang lebar, model SVM diharapkan dapat meningkatkan kemampuan generalisasi dan meminimalkan kesalahan klasifikasi(Avci dkk., 2023).

Keunggulan utama SVM terletak pada penggunaan fungsi kernel yang memungkinkan algoritma ini menangani data yang tidak dapat dipisahkan secara linear. Fungsi *kernel* seperti linear kernel, *polynomial kernel*, dan *radial basis function* (RBF) digunakan untuk memetakan data ke ruang berdimensi lebih tinggi sehingga pola *non-linear* dapat dipisahkan dengan lebih mudah. Fleksibilitas ini menjadikan SVM efektif pada dataset dengan dimensi tinggi, distribusi kompleks, maupun data dengan jumlah sampel relatif terbatas. Beberapa penelitian juga menunjukkan bahwa SVM memiliki performa yang stabil dan kompetitif dibandingkan metode klasifikasi lainnya, seperti *decision tree* dan *random forest*, terutama dalam aplikasi yang memerlukan presisi tinggi, misalnya klasifikasi penggunaan lahan dan tutupan lahan (LULC) berbasis citra satelit(Avci dkk., 2023).

Pendekatan *stylometry* telah banyak dimanfaatkan dalam tugas *authorship attribution* untuk mengenali penulis berdasarkan pola *linguistik* yang konsisten. *Stylometry* sendiri merupakan analisis statistik terhadap gaya bahasa dalam teks, mencakup fitur seperti panjang kata, distribusi kata fungsi, panjang kalimat, penggunaan tanda baca, kekayaan kosakata, hingga *n-grams*. Berbagai fitur *stylometry* (446 fitur, termasuk kata fungsi, panjang kata, *word endings*, tanda baca, dan pola sintaksis) diekstraksi dari korpus berupa 10 disertasi PhD yang kemudian dipotong ke dalam segmen 1.000, 5.000, dan 10.000 kata. Selanjutnya, metode pembelajaran mesin seperti *k-Nearest Neighbors* (k-NN) dan *Sequential Minimal Optimization* (SMO) untuk *Support Vector Machine* (SVM) digunakan untuk melakukan klasifikasi penulis. Hasil penelitian menunjukkan bahwa SVM (SMO) *consistently* mengungguli k-NN dalam semua skenario, terutama pada ukuran teks yang lebih kecil. Akurasi tertinggi mencapai 98% dengan menggunakan SVM dan segmen teks 10.000 kata, sedangkan kinerja k-NN cenderung lebih bervariasi dan menurun pada teks yang lebih pendek. Temuan ini menegaskan bahwa SVM efektif dalam menangkap pola gaya penulisan yang kompleks, serta lebih stabil terhadap variasi jumlah fitur dibandingkan k-NN.(Maurya dkk., 2021)

Penelitian lain menunjukkan bahwa keberhasilan *authorship attribution* sangat dipengaruhi oleh pemilihan fitur yang tepat dan proses prapemrosesan data yang baik. Tantangan seperti variasi gaya antar penulis serta keterbatasan himpunan data (dataset) menjadi isu yang perlu diperhatikan. Fitur-fitur *stylometry* dapat dibedakan menjadi tiga kategori utama, yaitu (1) fitur leksikal, seperti panjang kata dan distribusi kata fungsi; (2) fitur sintaktis, seperti penggunaan struktur kalimat; serta (3) fitur semantik, termasuk pemilihan kosakata dan gaya diskursif. Namun, tantangan signifikan tetap ada, antara lain variasi gaya antar penulis yang dapat tumpang tindih, serta keterbatasan dataset yang seringkali kecil atau tidak seimbang, sehingga menyulitkan model pembelajaran mesin untuk melakukan generalisasi yang baik. Masalah lain adalah keberadaan noise dalam teks, misalnya kesalahan ketik, penggunaan bahasa gaul, atau campuran bahasa, yang dapat memengaruhi ekstraksi fitur *stylometry*. Penelitian ini menegaskan bahwa dimasa depan perlu lebih berfokus pada strategi pemilihan fitur yang adaptif, teknik

prapemrosesan data yang lebih robust, serta pengembangan dataset yang lebih besar dan representatif, agar hasil atribusi penulis menjadi lebih akurat dan dapat diandalkan . (He dkk., 2024).

Studi terkini menunjukkan bahwa pendekatan pembelajaran mendalam (*deep learning*) telah membuka peluang baru dalam bidang *stylometry* dan *authorship attribution*. Model seperti *Convolutional Neural Networks* (CNN), *Recurrent Neural Networks* (RNN), serta arsitektur transformer mampu menangkap pola linguistik yang kompleks dan halus, melampaui keterbatasan metode statistik tradisional yang hanya mengandalkan fitur buatan seperti frekuensi kata, panjang kalimat, atau pola sintaksis. Keunggulan utama model berbasis deep learning terletak pada kemampuannya melakukan pembelajaran representasi secara hierarkis, sehingga lebih efektif dalam menganalisis teks pendek, multibahasa, maupun lintas-genre. Beberapa studi bahkan menunjukkan bahwa CNN dan RNN dapat mengungguli metode klasik seperti *n-grams* dan SVM dalam atribusi penulis, terutama pada skenario yang menantang dengan banyak penulis atau variasi gaya. Meskipun demikian, dalam penelian ini menekankan bahwa masih terdapat sejumlah tantangan, di antaranya adalah keterbatasan dalam interpretabilitas model, sehingga sulit menjelaskan alasan spesifik suatu teks diklasifikasikan ke penulis tertentu, serta kebutuhan akan dataset pelatihan yang besar dan seimbang agar model dapat melakukan generalisasi dengan baik. Hal ini menuntut pengembangan lebih lanjut dalam hal teknik interpretasi serta metode pembelajaran yang lebih efisien untuk data terbatas. (Sharma & Kumar, 2024)

Stylometry tidak hanya digunakan dalam identifikasi penulis, tetapi juga dalam analisis lain seperti estimasi tingkat kecerdasan (*intelligence quotient*/IQ). Pendekatan berbasis teknologi pencitraan saraf (*neuroimaging*) seperti *Magnetic Resonance Imaging* (MRI), *Electroencephalography* (EEG), dan *functional Near-Infrared Spectroscopy* (fNIRS) memiliki keterbatasan dari segi biaya dan aksesibilitas, sehingga muncul alternatif berbasis teks melalui analisis linguistik (Adebayo & Yampolskiy, 2022).

Dalam bidang deteksi plagiarisme, pendekatan *stylometry* juga memberikan kontribusi penting. Mengembangkan sebuah sistem berbasis *Support Vector*

Machine (SVM) dengan memanfaatkan kombinasi fitur leksikal, sintaksis, dan semantik untuk membedakan teks asli dan teks hasil plagiarisme. Fitur leksikal mencakup distribusi kata dan panjang kata, fitur sintaksis melibatkan pola struktur kalimat, sedangkan fitur semantik berfokus pada makna dan keterhubungan antar-kata dalam teks. Sistem ini dievaluasi menggunakan dataset standar PAN (*Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection*) 2013 dan 2014, yang memang dirancang khusus untuk kompetisi deteksi plagiarisme. Hasil pengujian menunjukkan bahwa metode berbasis *stylometry* dan SVM mampu mendeteksi berbagai bentuk plagiarisme, baik yang bersifat verbatim (penyalinan langsung) maupun paraphrased plagiarism (penulisan ulang dengan gaya berbeda). Dengan teknik seleksi fitur yang tepat, akurasi deteksi dapat ditingkatkan secara signifikan, sehingga sistem ini terbukti efektif dibandingkan metode konvensional berbasis pencocokan teks semata. (El-Rashidy dkk., 2024).

Deteksi plagiarisme merupakan salah satu bidang penelitian penting dalam pengolahan bahasa alami (*Natural Language Processing*) yang bertujuan untuk mengidentifikasi kesamaan antara dokumen. Plagiarisme sendiri dapat berbentuk verbatim plagiarism (penyalinan langsung), paraphrase plagiarism (penyusunan ulang kalimat dengan kata berbeda namun makna sama), hingga *translated plagiarism* (penerjemahan lintas bahasa). Oleh karena itu, metode pendeteksian plagiarisme harus mampu mengidentifikasi berbagai jenis kesamaan teks dengan akurasi tinggi.

Penelitian lain juga membahas perbandingan tingkat akurasi dari berbagai metode deteksi plagiarisme, di antaranya *Cosine Similarity*, *Jaccard Coefficient*, dan Rabin-Karp. Metode *Cosine Similarity* bekerja dengan mengukur kesamaan sudut vektor antar dokumen berdasarkan representasi kata, sehingga efektif dalam menemukan kesamaan konteks. *Jaccard Coefficient* digunakan untuk menghitung irisan dan gabungan kata antar dokumen, namun kinerjanya cenderung menurun ketika teks memiliki banyak variasi sinonim. Sedangkan *Rabin-Karp* merupakan algoritma pencocokan *string* berbasis *hashing* yang lebih sesuai untuk deteksi plagiarisme verbatim. Hasil evaluasi menunjukkan bahwa tingkat akurasi metode sangat dipengaruhi oleh jenis plagiarisme yang diuji. Misalnya, *Cosine Similarity*

lebih unggul dalam mendeteksi plagiarisme berbasis parafrasa karena mampu mempertahankan informasi semantik, sementara *Rabin-Karp* lebih tepat digunakan untuk deteksi penyalinan langsung (*copy-paste*). Hal ini menegaskan bahwa tidak ada satu metode tunggal yang paling efektif untuk semua jenis plagiarisme, sehingga pengembangan sistem deteksi modern biasanya mengombinasikan beberapa pendekatan untuk memperoleh hasil yang lebih akurat dan komprehensif(Rahma & Taufiq, 2024).

Dari berbagai studi di atas, dapat disimpulkan bahwa pendekatan stylometry tidak hanya efektif dalam mengidentifikasi penulis, tetapi juga dalam pengembangan sistem deteksi plagiarisme dan analisis karakteristik penulis lainnya. Penelitian ini menggabungkan *stylometry* dengan algoritma *Support Vector Machine* (SVM) untuk *authorship attribution*, kemudian dilanjutkan dengan deteksi plagiarisme berbasis kemiripan gaya penulisan dan semantik menggunakan *Sentence*-BERT (*Bidirectional Encoder Representations from Transformers*).

2.2 Dasar Teori

2.2.1 Deteksi Plagiarisme

Plagiarisme merupakan tindakan mengambil atau menjiplak karya orang lain baik berupa ide, gagasan, maupun tulisan tanpa menyebutkan sumber aslinya dan mengakuinya sebagai karya sendiri. Menurut Permendiknas No. 17 Tahun 2010, plagiarisme adalah perbuatan yang disengaja atau tidak disengaja untuk mendapatkan kredit akademik dengan mengutip sebagian atau seluruh karya ilmiah orang lain tanpa mencantumkan sumber secara tepat.

Seiring berkembangnya teknologi dan kemudahan akses informasi melalui internet, praktik plagiarisme semakin marak terjadi, terutama di lingkungan akademik. Hal ini berdampak pada menurunnya integritas akademik serta mencederai nilai-nilai kejujuran dan etika ilmiah. Oleh karena itu, deteksi plagiarisme menjadi langkah penting dalam menjaga kualitas dan orisinalitas karya ilmiah. Deteksi plagiarisme dilakukan dengan membandingkan kesamaan kata, kalimat, atau struktur teks antara karya yang diuji dengan sumber referensi lain. Salah satu aplikasi populer yang digunakan adalah Turnitin, sebuah perangkat lunak berbasis web yang mampu mengukur tingkat kemiripan teks (*similarity index*)

secara global terhadap berbagai sumber daring seperti jurnal, artikel, dan *repository* institusi. Penerapan sistem deteksi plagiarisme secara otomatis melalui perangkat lunak seperti Turnitin tidak hanya memudahkan proses verifikasi, tetapi juga mendorong peningkatan integritas akademik, yakni kejujuran, tanggung jawab, dan orisinalitas dalam penulisan karya ilmiah. Strategi ini penting dalam membangun budaya akademik yang sehat dan bebas dari praktik penjiplakan (Silalahi dkk., 2024).

2.2.2 Novel Sastra

Sastra merupakan bentuk ekspresi artistik yang menggunakan bahasa sebagai medium utamanya untuk menyampaikan pengalaman, pemikiran, dan perasaan manusia. Sastra adalah hasil kreativitas pengarang yang bersumber dari kehidupan manusia secara langsung melalui rekaan dengan bahasa sebagai mediumnya. Novel, sebagai salah satu bentuk karya sastra, adalah karya fiksi naratif yang menyuguhkan tokoh-tokoh serta serangkaian peristiwa dalam alur yang terstruktur. Novel memuat keseluruhan unsur kehidupan yang dapat bersifat nyata ataupun rekaan, dan karena itu sangat erat hubungannya dengan imajinasi pengarang (Saragih dkk., 2021).

Dalam penelitian ini, novel sastra dipilih sebagai data penelitian karena memiliki beberapa keunggulan, antara lain:

1. Kekayaan Bahasa

   Novel mengandung variasi kosakata yang luas, penggunaan gaya bahasa yang beragam, serta kompleksitas kalimat yang tinggi. Hal ini memungkinkan sistem untuk menganalisis *stylometric features* seperti panjang kalimat, keragaman kosakata, maupun distribusi tanda baca.

2. Ciri Khas Penulisan

   Setiap penulis pasti memiliki gaya tulisan yang khas (*linguistic fingerprint*). Misalnya Mark Twain dikenal dengan penggunaan bahasa sehari-hari dan humor satir, sementara Mary Shelley lebih cenderung pada deskripsi romantik dan metaforis. Perbedaan gaya inilah yang dapat dianalisis dengan *stylometry*.

3. Panjang Dokumen

Novel sastra mempunya ribuan hingga ratusan ribu kata, sehingga teks novel cocok untuk metode *chunking*. Teks yang Panjang ini nanti dapat dipecah menjadi beberapa potongan berukuran 1000, 5000, atau 10000 kata untuk dianalisis. Hal ini memungkinkan sistem bekerja lebih efisien dan tetap mempertahankan ciri khas penulis.

4. Relevansi untuk Deteksi Plagiarisme

Plagiarisme tidak hanya terjadi pada teks akademik, tetapi juga dalam karya sastra seperti novel. Dengan menggunakan novel sastra, sistem dapat menguji sejauh mana metode *stylometry* dan *semantic similarity* mampu mengenali persamaan atau perbedaan antar teks dalam ranah kreatif.

2.2.3 Preprocessing

Preprocessing atau pra-pemrosesan data merupakan tahap awal yang sangat penting dalam analisis data, termasuk dalam bidang *text mining, machine learning*, maupun *natural language processing* (NLP). Pra-pemrosesan bertujuan untuk meningkatkan kualitas data sehingga lebih siap digunakan dalam tahap analisis dan pemodelan. Data mentah yang diperoleh biasanya mengandung noise, inkonsistensi, duplikasi, maupun informasi yang tidak relevan, sehingga diperlukan transformasi agar representasi data menjadi lebih bersih dan bermakna (Çetin & Yıldız, 2022).

Secara umum, *preprocessing* melibatkan beberapa teknik utama, yaitu:

1. Data *Cleaning* (Pembersihan Data)

Tahap ini menghapus data yang hilang, duplikat, atau tidak konsisten. Dalam teks, proses ini mencakup penghapusan tanda baca yang tidak relevan, angka, simbol khusus, dan *typo correction*.

2. Data Integration (Integrasi Data)

Penggabungan data dari berbagai sumber agar membentuk dataset yang lebih lengkap dan konsisten. Misalnya, dalam analisis *stylometry*, teks dari berbagai dokumen atau penulis digabungkan ke dalam satu korpus.

3. Data Transformation (Transformasi Data)

Melibatkan normalisasi, stemming, dan lemmatisasi agar bentuk kata seragam. Contohnya, kata "*running*", "*runs*", dan "*ran*" diubah menjadi bentuk dasar "*run*".

4. Data *Reduction* (Reduksi Data)

   Mengurangi dimensi atau jumlah fitur tanpa menghilangkan informasi penting, misalnya dengan *feature selection* atau *dimensionality reduction* (PCA). Dalam *stylometry*, ini membantu mengurangi beban komputasi.

5. Data *Discretization* (Diskretisasi Data)

   Mengubah data kontinu menjadi data kategorikal. Misalnya, panjang kalimat bisa dikategorikan menjadi pendek, sedang, dan panjang.

Dalam konteks analisis teks dan *stylometry*, *preprocessing* biasanya mencakup tahapan seperti *tokenization*, *stopword removal*, *stemming*/*lemmatization*, serta normalisasi teks (huruf kecil, penghapusan spasi ganda, dan lain-lain). Tahap ini sangat krusial karena kualitas fitur yang diekstraksi sangat tergantung pada kualitas data hasil pra-pemrosesan.

2.2.4 Chunking

Chunking merupakan strategi yang digunakan untuk membagi teks panjang menjadi potongan-potongan kecil (*chunks*) agar lebih mudah dipahami dan dianalisis. Strategi ini terbukti efektif dalam meningkatkan pemahaman bacaan karena membantu pembaca fokus pada bagian-bagian penting dari teks tanpa merasa terbebani oleh keseluruhan isi. Penerapan *chunking* dalam pembelajaran membaca terbukti meningkatkan kemampuan siswa memahami teks berbahasa Inggris secara signifikan. Dalam konteks penelitian ini, *chunking* digunakan untuk membagi teks novel menjadi bagian-bagian berukuran seragam (misalnya 1000 kata) sebagai dasar dalam proses ekstraksi fitur *stylometry* dan analisis kemiripan teks untuk *authorship attribution* dan deteksi plagiarisme (RISAKOTTA, 2023).

Dalam penelitian ini, teknik *chunking* memiliki beberapa tujuan:

1. Normalisasi panjang teks

   Setiap potongan teks memiliki panjang relatif sama sehingga ekstraksi fitur lebih adil dan tidak bias.

2. Mengatasi keterbatasan model

Model *Sentence*-BERT (SBERT) hanya dapat menerima jumlah token terbatas (512–1024 token), sehingga teks panjang harus dipotong.

3. Efisiensi komputasi

   Analisis per *chunk* lebih ringan dibandingkan memproses keseluruhan dokumen.

4. Deteksi plagiarisme local

   Chunking memungkinkan pelacakan bagian teks tertentu yang terindikasi memiliki kemiripan dengan karya lain, bukan hanya menilai keseluruhan dokumen.

Dalam penelitian ini, *chunking* berperan penting sebagai strategi metodologis untuk memastikan proses analisis berjalan optimal. Pada analisis *semantic similarity*, penggunaan *chunk* berukuran kecil membantu menjaga konteks kalimat tetap utuh sekaligus menyesuaikan dengan batas kapasitas model seperti SBERT yang hanya dapat memproses jumlah token terbatas. Sementara itu, pada analisis *stylometry*, penggunaan *chunk* berukuran besar lebih stabil dalam menangkap ciri khas penulis, misalnya *average sentence length, type-token ratio*, maupun *hapax legomena ratio*, karena semakin banyak teks yang dianalisis semakin representatif pola gaya bahasanya. Dengan demikian, *chunking* tidak hanya berfungsi sebagai teknik pemotongan teks panjang, tetapi juga sebagai langkah penting untuk meningkatkan akurasi, efisiensi, dan ketepatan hasil deteksi plagiarisme berbasis makna maupun gaya bahasa.

2.2.5 Stylometry

Stylometry adalah cabang ilmu yang mempelajari gaya penulisan seseorang menggunakan pendekatan statistik dan linguistik. Analisis ini mencakup fitur seperti frekuensi kata, panjang kalimat, struktur sintaksis, dan pilihan kosakata. *Stylometry* telah banyak diterapkan dalam bidang identifikasi penulis (*authorship attribution*), pemetaan karakteristik penulis (*author profiling*), dan forensik linguistik. Teknik ini terbukti efektif dalam mengidentifikasi ciri demografis seperti usia, jenis kelamin, bahasa ibu, dan kepribadian penulis.

Dalam konteks estimasi tingkat kecerdasan (*intelligence quotient*/IQ), *stylometry* digunakan untuk menilai kualitas dan kompleksitas teks yang

diasumsikan berkorelasi dengan kecerdasan penulis. Beberapa fitur utama yang digunakan meliputi *lexical diversity* (keragaman kosakata), *syntactic complexity* (kompleksitas sintaksis), dan *collegiate word ratio* (rasio kata tingkat perguruan tinggi) (Adebayo & Yampolskiy, 2022).

Stylometry adalah cabang dari linguistik komputasional yang menganalisis gaya penulisan seseorang menggunakan fitur-fitur statistik dari teks. Dalam konteks *authorship attribution, stylometry* digunakan untuk mengekstrak ciri-ciri khas penulisan, seperti frekuensi kata, panjang kalimat, atau distribusi kata-kata fungsional (*function words*). *Stylometry* terbukti efektif dalam membedakan antara penulis yang berbeda bahkan pada topik yang berbeda (*cross-topic attribution*) (Sarwar dkk., 2024).

Dalam penelitian ini, *stylometry* diimplementasikan melalui ekstraksi fitur-fitur numerik dari teks sehingga teks dapat direpresentasikan dalam bentuk vektor. Vektor tersebut kemudian dianalisis menggunakan algoritma pembelajaran mesin (misalnya *Support Vector Machine*) untuk membedakan gaya antar penulis. Adapun sepuluh fitur utama yang digunakan dalam penelitian ini adalah:

1. *Average Sentence Length*

   Menghitung panjang rata-rata kalimat dalam satu chunk, dihitung berdasarkan jumlah kata per kalimat. Menggambarkan kecenderungan penulis dalam menyusun kalimat panjang atau pendek.

2. *Average Word Length*

   Mengukur rata-rata panjang kata dalam chunk (jumlah huruf per kata) dan mewakili tingkat kompleksitas kosakata yang digunakan.

3. *Type-Token Ratio* (TTR)

   Perbandingan antara jumlah kata unik dan jumlah total kata.

$$\text{TTR} = \frac{jumlah\ kata\ unik}{jumlah\ total\ kata} \tag{1.1}$$

   Rumus TTR (*Type-Token Ratio*) mengukur variasi kosakata dalam teks dengan membandingkan jumlah kata unik (*type*) terhadap jumlah total kata (token). Kata unik adalah kata yang tidak berulang, sementara total kata mencakup semua kata termasuk yang berulang. Nilai TTR berkisar antara 0 hingga 1; semakin tinggi nilainya, semakin bervariasi kosakata dalam teks

tersebut. Rumus ini sering digunakan dalam analisis linguistik dan stylometry untuk menilai kekayaan atau kompleksitas bahasa dalam suatu tulisan.

4. *Hapax Legomena Ratio*

   Rasio jumlah kata yang hanya muncul sekali terhadap total kata. Fitur ini menggambarkan kekayaan kosakata seorang penulis. *Hapax Legomena Ratio* digunakan sebagai salah satu ciri khas (*fingerprint*) penulis. Nilai rasio ini, bila dikombinasikan dengan fitur lain seperti *type-token ratio* atau *average word length*, mampu membantu sistem membedakan gaya penulisan antar penulis.

5. *Vowel Ratio*

   Proporsi huruf vokal (a, e, i, o, u) terhadap seluruh huruf alfabet dalam teks. Mengindikasikan pola fonetik yang khas dari gaya penulisan.

6. *Stopword Ratio*

   Persentase kata umum (*stopwords*) terhadap total kata, berdasarkan daftar dari NLTK. Menunjukkan seberapa fungsional atau padat makna suatu teks.

7. *Punctuation Ratio*

   Rasio jumlah tanda baca (seperti tanda titik, koma dan lainnya) terhadap total karakter. Menunjukkan kebiasaan penggunaan tanda baca yang unik pada setiap penulis.

8. *Character n-gram Entropy*

   Mengukur kompleksitas distribusi karakter dalam chunk, berdasarkan entropi dari *n-gram* karakter (*default* n=3).

$$H = -\sum P_i \ log_2 P_i \tag{1.2}$$

   Rumus tersebut digunakan untuk menghitung entropy, yaitu ukuran ketidakpastian atau keragaman distribusi dalam suatu himpunan data, termasuk teks. Dalam rumus ini, H merepresentasikan nilai entropy secara keseluruhan. Simbol $\sum$ menunjukkan proses penjumlahan untuk seluruh elemen yang ada, yaitu semua jenis kata atau simbol yang dianalisis. P_i adalah probabilitas kemunculan elemen atau kata ke-i dalam teks, sedangkan $log_2 P_i$ merupakan logaritma basis 2 dari probabilitas tersebut. Perkalian antara P_i dan $log_2 P_i$ menghitung kontribusi setiap elemen terhadap total entropy,

dan tanda negatif di depan penjumlahan memastikan bahwa hasilnya positif, karena logaritma dari bilangan pecahan adalah negatif. Semakin tinggi nilai entropy, semakin besar variasi atau ketidakpastian dalam distribusi kata, yang mencerminkan keragaman bahasa dalam teks tersebut.

9. *POS Tag Distribution*

   Rasio penggunaan kategori kata seperti kata benda (*NOUN*), kata kerja (VERB), dll. Mewakili gaya struktural penulis dalam menyusun kalimat.

10. *Chunk Length*

    Jumlah total kata dalam *chunk* (sebagai fitur tambahan untuk kontrol). Berguna untuk memastikan setiap potongan data memiliki skala yang sebanding.

Seiring dengan perkembangan teknologi, *stylometry* kini banyak dipadukan dengan machine learning dan deep learning untuk meningkatkan akurasi analisis. Studi terbaru menunjukkan bahwa metode *support vector machine* (SVM), *decision tree*, maupun *neural networks* dapat digunakan untuk klasifikasi penulis berdasarkan fitur *stylometrik* (Zheng et al., 2022). Selain itu, dengan hadirnya *contextual embeddings* berbasis *transformer*, analisis *stylometry* tidak hanya terbatas pada pola permukaan (*surface features*), tetapi juga dapat menangkap aspek semantik yang lebih dalam.

Stylometry memiliki peran penting dalam berbagai aplikasi praktis, seperti *authorship attribution*, deteksi plagiarisme, analisis forensik digital, serta kajian sejarah sastra. Dengan pendekatan ini, teks bukan hanya dipandang sebagai medium komunikasi, tetapi juga sebagai representasi gaya personal penulis yang dapat dianalisis secara ilmiah(Gorman, 2024).

2.2.6 Authorship Attribution

Authorship attribution (atribusi kepenulisan) adalah proses menentukan penulis asli dari suatu teks anonim dari sekumpulan penulis potensial. Teknik ini memiliki berbagai aplikasi penting di dunia nyata, termasuk dalam deteksi plagiarisme, forensik digital, identifikasi ancaman keamanan, dan analisis perilaku media sosial. Tujuan utama dari *authorship attribution* adalah mengungkap gaya penulisan unik yang konsisten dimiliki oleh setiap individu (Sarwar dkk., 2024).

Authorship attribution, juga dikenal sebagai *author recognition* atau *author verification*, merupakan proses untuk mengidentifikasi penulis suatu dokumen berdasarkan gaya penulisannya. Proses ini telah digunakan dalam berbagai konteks, mulai dari identifikasi penulis karya sastra klasik seperti *Federalist Papers* hingga deteksi pelaku dalam kasus kejahatan siber dan penipuan akademi. Tujuan utama dari *authorship attribution* adalah untuk menangkap pola atau ciri khas dalam gaya penulisan yang membedakan satu penulis dari yang lain (Maurya et al., 2021).

Dalam penelitian ini, teknik *Authorship Attribution* (AA) diterapkan melalui dua pendekatan utama, yaitu *stylometry* dan *semantic similarity*. Pendekatan *stylometry* menitikberatkan pada ekstraksi sepuluh fitur linguistik seperti *average sentence length, type-token ratio, hapax legomena ratio,* distribusi kata fungsi, hingga *entropi n-gram* karakter. Fitur-fitur ini digunakan untuk menangkap gaya tulis penulis yang biasanya konsisten meskipun topik atau kosakata berubah, kemudian dianalisis menggunakan algoritma *Support Vector Machine* (SVM) untuk memprediksi kesamaan penulis antar teks. Sementara itu, pendekatan *semantic* similarity berfokus pada representasi makna teks, dengan memanfaatkan metode TF-IDF sebagai representasi sederhana dan *Sentence*-BERT (SBERT) sebagai representasi semantik yang lebih mendalam, guna mengukur kedekatan makna antar potongan teks (*chunks*).

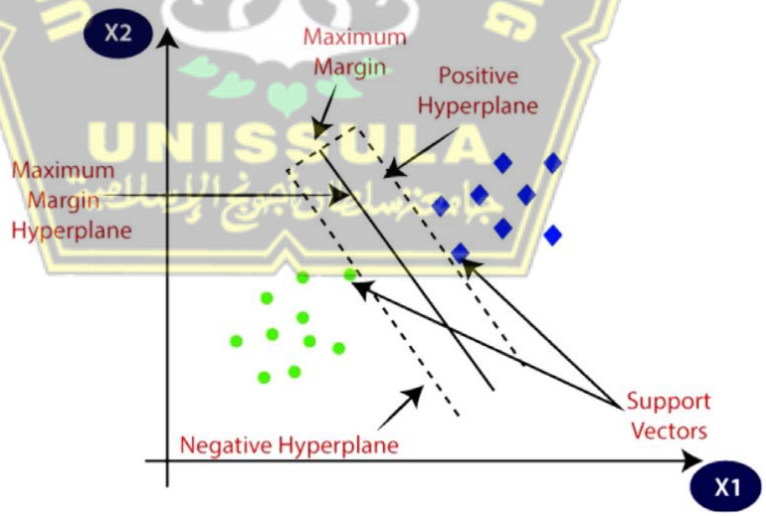
Kedua pendekatan ini saling melengkapi, *stylometry* lebih andal dalam mengidentifikasi pola gaya tulis sebagai penulis, sedangkan *semantic similarity* unggul dalam mendeteksi kesamaan konten meskipun telah mengalami parafrasa atau modifikasi. Dengan kombinasi keduanya, sistem AA dalam penelitian ini tidak hanya mengandalkan persamaan kata demi kata, tetapi juga mampu mengungkap kesamaan dari sisi gaya bahasa maupun makna, sehingga meningkatkan keandalan deteksi plagiarisme secara lebih komprehensif.

Perkembangan terbaru dalam atribusi kepenulisan semakin banyak memanfaatkan pendekatan *machine learning* dan *deep learning* yang mampu mengekstrak pola lebih kompleks dari teks dalam skala besar. Pada awalnya, metode sederhana seperti teknik kompresi teks digunakan untuk membedakan penulis, namun kemudian berkembang dengan adopsi algoritma seperti *support*

vector machine (SVM), *random forest*, hingga *neural networks* untuk meningkatkan akurasi prediksi. Dengan hadirnya model bahasa berbasis transformer, atribusi kepenulisan kini dapat dilakukan dengan memanfaatkan *contextual embeddings* yang mampu menangkap representasi semantik dan sintaksis secara lebih mendalam. Pendekatan ini terbukti efektif untuk mengidentifikasi ciri khas penulis, bahkan pada teks yang relatif singkat atau telah mengalami parafrasa, sehingga relevan untuk diterapkan dalam bidang keamanan informasi, deteksi plagiarisme, maupun investigasi forensik digital.(Assael dkk., 2022).

2.2.7 Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah algoritma pembelajaran mesin yang digunakan untuk tugas klasifikasi dan regresi, terutama dalam kategori *supervised learning*. Algoritma ini dikembangkan oleh Vladimir Vapnik dan Alexey Chervonenkis pada tahun 1963, dan menjadi sangat populer karena kemampuannya dalam menangani *dataset* berdimensi tinggi dan memberikan batas klasifikasi yang optimal (Bansal dkk., 2022).



Gambar 2. 1 Ilustrasi *Support Vector Machine* (Bansal dkk., 2022)

Garis tengah pada ilustrasi SVM merepresentasikan *hyperplane*, yaitu batas keputusan optimal yang digunakan untuk memisahkan dua kelas data. Di kedua sisi *hyperplane* terdapat dua garis sejajar yang menggambarkan

margin, yaitu jarak terdekat dari *hyperplane* ke data masing-masing kelas. Titik-titik yang berada di dekat atau tepat pada margin disebut sebagai *support vectors*, karena posisi merekalah yang secara langsung menentukan letak dan orientasi *hyperplane*, sehingga berperan penting dalam proses klasifikasi yang optimal.

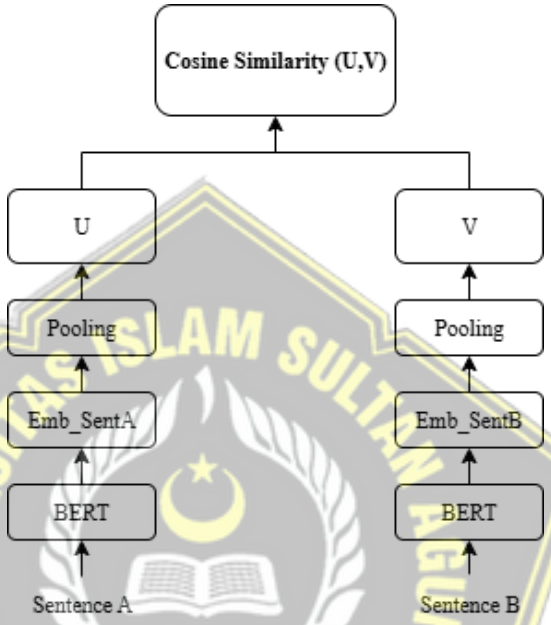
Support Vector Machine (SVM) merupakan salah satu algoritma machine learning berbasis pembelajaran terawasi (*supervised learning*) yang digunakan secara luas untuk klasifikasi dan regresi. SVM bekerja dengan mencari sebuah *hyperplane* optimal yang memisahkan data ke dalam dua atau lebih kelas dengan margin maksimum. Konsep utama SVM adalah memilih garis pemisah (pada data dua dimensi) atau bidang pemisah (pada dimensi lebih tinggi) yang tidak hanya mampu memisahkan kelas, tetapi juga memberikan jarak maksimum terhadap data terdekat dari masing-masing kelas, yang disebut *support vectors*. Dengan cara ini, SVM berusaha meminimalkan kesalahan klasifikasi dan meningkatkan kemampuan generalisasi model.

Selain itu, SVM memiliki fleksibilitas tinggi karena dapat menggunakan berbagai fungsi *kernel* untuk menangani data non-linear. Fungsi kernel, seperti *linear kernel*, *polynomial kernel*, atau *radial basis function* (RBF), memungkinkan data yang tidak terpisahkan secara linear di ruang asli dipetakan ke dalam ruang berdimensi lebih tinggi sehingga menjadi lebih mudah dipisahkan. Karakteristik ini menjadikan SVM unggul dalam menangani masalah klasifikasi yang kompleks, termasuk pada data dengan dimensi tinggi, distribusi yang tidak seragam, maupun pola yang tidak linier (Avci dkk., 2023).

2.2.8 Sentence-BERT (SBERT)

SBERT merupakan pengembangan dari model BERT (*Bidirectional Encoder Representations from Transformers*) yang dirancang khusus untuk menghasilkan representasi kalimat (*sentence embeddings*) yang bermakna secara semantik. SBERT mengintegrasikan arsitektur *Siamese Network* dengan BERT dan

menambahkan proses pooling agar dapat menghasilkan vektor tetap dari kalimat dengan panjang bervariasi. SBERT sangat cocok digunakan untuk menghitung kesamaan antar kalimat (*semantic textual similarity*), *semantic search*, dan *paraphrase mining*, karena kemampuannya menghasilkan *embedding* yang bisa dibandingkan menggunakan *cosine similarity* (Santander-Cruz dkk., 2022).



Gambar 2. 2 SBERT *architecture* (Santander-Cruz dkk., 2022)

Gambar 2.2 menampilkan dua input kalimat yang dimasukkan ke model BERT paralel (*Siamese*), lalu dilakukan *pooling* untuk menghasilkan dua *embedding* kalimat (u dan v), yang kemudian dibandingkan menggunakan *cosine similarity*.

Cosine Similarity digunakan untuk mengukur tingkat kesamaan semantik antara dua kalimat atau antara deskripsi pasien dengan ide utama atau ground-truth dalam tes narasi (seperti *cookie thief test*). Ini menjadi bagian penting dalam proses klasifikasi karena memungkinkan model untuk menilai seberapa relevan narasi pasien terhadap informasi yang seharusnya disampaikan(Santander-Cruz dkk., 2022).

$$Sim_{cos}(d,q) = \frac{\sum \left(P(n,d) \times P(n,q)\right)}{\sqrt{\sum P(n,d)^2} \times \sqrt{P(n,q)^2}} \quad (1.3)$$

Dalam rumus *cosine similarity* tersebut, setiap simbol memiliki fungsi yang spesifik untuk menghitung tingkat kemiripan antara dua teks, yaitu dokumen (*d*)

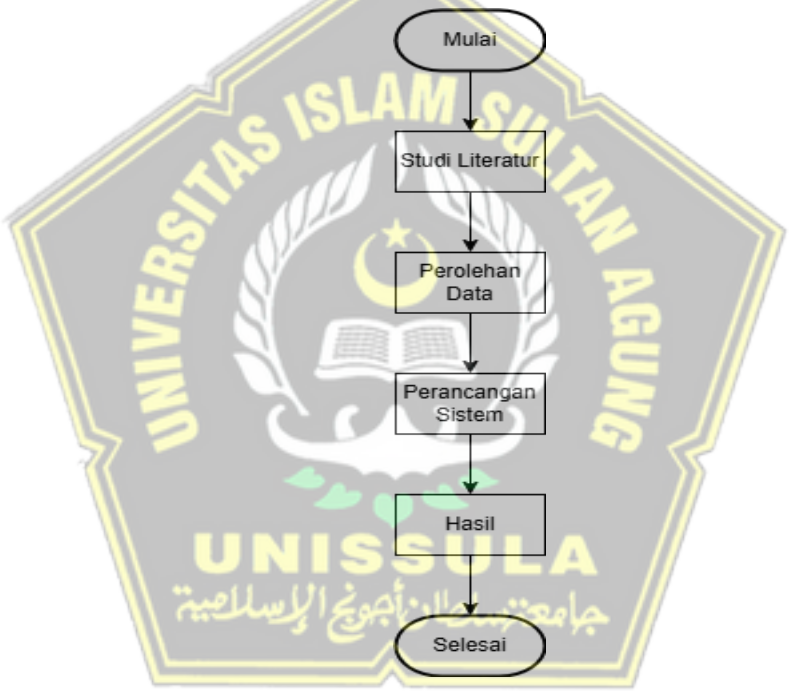
dan kueri (q). Simbol $Sim_{cos}(d, q)$ menunjukkan nilai *cosine similarity* antara dokumen d dan kueri q. Fungsi $P(n, d)$ merepresentasikan bobot kata ke-n dalam dokumen d, sedangkan $P(n, q)$ adalah bobot kata ke-n dalam kueri q, yang biasanya diperoleh dari teknik representasi seperti TF-IDF atau *word embedding*. Tanda $\sum$ pada pembilang mengindikasikan penjumlahan dari hasil kali antara bobot kata yang sama dalam dokumen dan kueri, menunjukkan sejauh mana kedua teks tersebut memiliki kesamaan dalam penggunaan kata. Di penyebut, terdapat akar kuadrat dari penjumlahan kuadrat bobot kata dalam dokumen $\sqrt{\sum P(n, d)^2}$ dan kueri $\sqrt{\sum P(n, q)^2}$, yang masing-masing menghitung panjang (norma) vektor dari dokumen dan kueri. Dengan membagi hasil *dot product* di pembilang dengan hasil kali norma kedua vektor di penyebut, rumus ini menghasilkan nilai kemiripan yang ternormalisasi antara 0 dan 1, di mana nilai mendekati 1 menandakan bahwa dokumen dan kueri sangat mirip secara semantik.

Dalam penelitian ini, proses atribusi kepenulisan dan deteksi plagiarisme berbasis teks dilakukan melalui pendekatan kuantitatif yang menghasilkan nilai representasi dari teks. Representasi ini penting karena teks pada dasarnya bersifat kualitatif, sehingga perlu diubah menjadi bentuk numerik agar dapat diproses oleh algoritma machine learning maupun deep learning.

Dua jenis representasi yang digunakan adalah *stylometry* dan *semantic similarity*. *Stylometry* berfokus pada pengukuran ciri khas gaya bahasa penulis, misalnya panjang rata-rata kalimat, variasi kosakata, hingga distribusi kata fungsi. Ciri-ciri ini relatif stabil dan menjadi "sidik jari" penulis, sehingga dapat digunakan untuk mengidentifikasi atau membedakan penulis antar teks. Sementara itu, *semantic similarity* berfokus pada kesamaan makna atau isi teks, bukan sekadar bentuk katanya. Representasi ini memungkinkan sistem mendeteksi kemiripan meskipun terdapat parafrasa atau perbedaan pilihan kata. Dengan mengombinasikan keduanya, sistem dapat menilai teks dari dua sisi yaitu gaya tulis dan isi makna. Hal ini membuat hasil analisis lebih akurat serta relevan untuk atribusi kepenulisan maupun deteksi plagiarisme.

BAB III
METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen komputasional. Tujuan penelitian adalah menguji efektivitas metode *stylometry* dan *machine learning* dalam proses identifikasi penulis (*authorship attribution*) serta deteksi plagiarisme pada teks novel berbahasa Inggris. Teknik yang digunakan bersifat eksploratif dan deskriptif, dengan fokus pada analisis kemiripan teks dan pengukuran akurasi prediksi penulis. Secara umum, alur penelitian dapat dijelaskan sebagai berikut:



Gambar 3. 1 Alur Penelitian

3.1 Studi Literatur

Dalam penelitian ini, penulis merujuk pada berbagai sumber referensi seperti jurnal ilmiah, makalah konferensi, artikel, dan skripsi terdahulu, serta memperluas wawasan melalui berbagai situs web resmi dan literatur digital. Tinjauan ini dilakukan dengan tujuan untuk mempelajari teori-teori yang relevan sebagai dasar dalam menganalisis dan memahami konsep sistem yang digunakan, pengolahan data teks, teknik ekstraksi fitur *stylometry*, serta metode klasifikasi menggunakan

algoritma *Support Vector Machine* (SVM) dan model pemetaan semantik berbasis *Sentence*-BERT (SBERT).

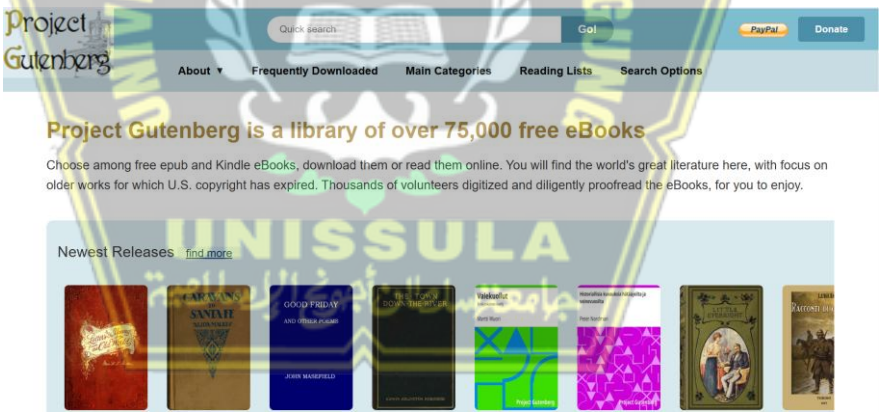
3.2 Perolehan Data

3.2.1 Sumber Data

Data yang digunakan dalam penelitian ini terdiri dari dua jenis utama, yaitu data teks asli dan data teks tiruan novel penulis asli, yang keduanya digunakan untuk proses *authorship attribution* dan deteksi plagiarisme berbasis *stylometry* dan pembelajaran mesin.

1) Data Teks Asli

Data ini diperoleh dari situs *Project Gutenberg* (*https://www.gutenberg.org*), sebuah repositori terbuka yang menyediakan berbagai karya sastra klasik dalam domain publik. Dalam penelitian ini, dipilih beberapa penulis berbahasa Inggris yang berbeda genre dan memiliki gaya menulis yang khas. Dari masing-masing penulis dipilih dua hingga tiga novel, yang kemudian dijadikan sebagai korpus utama.



Gambar 3. 2 Situs *Project Gutenberg*

2) Data Teks Tiruan

Data ini merupakan teks yang dihasilkan oleh model bahasa seperti GPT, yang dirancang untuk meniru gaya penulisan masing-masing penulis berdasarkan data teks asli. Tujuan utama dari data ini adalah untuk digunakan dalam pengujian sistem deteksi plagiarisme berbasis kemiripan semantik (menggunakan *Sentence*-BERT) dan kemiripan gaya (*stylometry*). Teks buatan ini merepresentasikan skenario realistis plagiarisme modern, di mana

kecerdasan buatan dimanfaatkan untuk menghasilkan tulisan yang menyerupai karya sastra asli, baik dari segi makna maupun gaya bahasa. Dengan membandingkan potongan teks buatan terhadap potongan teks asli dari penulis yang sama, sistem dapat mengidentifikasi indikasi plagiarisme melalui dua pendekatan, yaitu kesamaan semantik dan kesamaan gaya.

3.2.2 Format Data

Data yang diperoleh disimpan dalam format plain text (.txt) untuk memudahkan proses pembacaan, *preprocessing*, serta pemotongan teks (*chunking*). Format teks polos dipilih karena bebas dari elemen tambahan seperti *layout*, gambar, atau metadata, sehingga analisis dapat difokuskan pada konten linguistik. Setiap file teks mewakili satu dokumen utuh yang kemudian diproses lebih lanjut pada tahap preprocessing dan ekstraksi fitur.

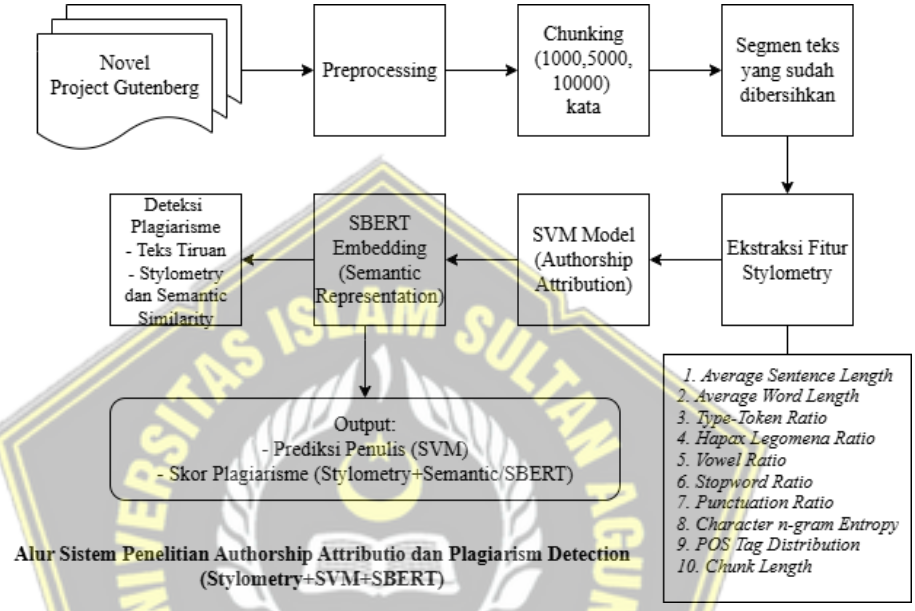
3.2.3 Teknik Pengumpulan Data

Pengumpulan data dalam penelitian ini dilakukan melalui dua sumber utama. Pertama, data asli diperoleh dari situs Project Gutenberg, yaitu kumpulan novel sastra dalam domain publik yang dipilih karena mewakili gaya penulisan khas masing-masing pengarang. Pemilihan Gutenberg sebagai sumber data didasarkan pada legalitas dan kelengkapan koleksi naskah digital yang tersedia.

Kedua, penelitian ini juga menggunakan data uji khusus yang merupakan hasil tiruan teks yang dihasilkan oleh model bahasa GPT. Teks tiruan ini dibuat dengan tetap menyesuaikan gaya tulis dari penulis asli yang terdapat pada data utama, sehingga menghasilkan variasi yang menyerupai karya aslinya. Tujuan dari pembuatan data uji khusus ini adalah untuk mensimulasikan kasus plagiarisme, baik melalui parafrasa maupun manipulasi struktur kalimat, sehingga sistem dapat dievaluasi dalam mendeteksi kesamaan berdasarkan gaya penulisan (*stylometry*) maupun kemiripan makna (*semantic similarity*). Dengan strategi ini, dataset penelitian tidak hanya kaya secara variasi, tetapi juga relevan untuk menguji keandalan sistem dalam skenario nyata deteksi plagiarisme.

3.3 Perancangan Sistem

Perancangan sistem ini bertujuan untuk menggambarkan alur kerja metode yang digunakan dalam penelitian, mulai dari proses pengambilan data hingga menghasilkan keluaran berupa prediksi penulis dan skor plagiarisme. Sistem ini terdiri dari beberapa tahapan utama sebagai berikut:



Gambar 3. 3 Perancangan Sistem *Authorship Attribution* dan Deteksi Plagiarisme

Data penelitian ini dikumpulkan dari *Project Gutenberg* yang terdiri atas 15 novel karya lima penulis berbeda (masing-masing penulis memiliki tiga novel). Pemilihan lima penulis dengan jumlah novel yang seimbang bertujuan agar distribusi data tetap proporsional sekaligus menghadirkan variasi gaya bahasa yang cukup beragam, sehingga proses *authorship attribution* dan deteksi plagiarisme dapat diuji secara lebih representatif.

3.3.1 Preprocessing Teks

Tahap awal sistem adalah *preprocessing* teks, yaitu proses pembersihan data agar siap untuk dianalisis. Proses ini mencakup penghapusan tanda baca, angka, karakter khusus, serta normalisasi huruf menjadi huruf kecil. Selain itu, dilakukan juga tokenisasi untuk memisahkan teks ke dalam unit kata atau kalimat. Tahap *preprocessing* sangat penting karena hasil ekstraksi fitur maupun representasi semantik sangat bergantung pada kualitas teks yang telah dibersihkan.

3.3.2 Teknik Chunking

Setelah teks diproses, langkah selanjutnya adalah membagi teks menjadi segmen (*chunks*) dengan ukuran tertentu. Proses *chunking* dilakukan pada tiga skala berbeda, yaitu 1000, 5000, dan 10000 kata, dengan tujuan untuk membagi novel menjadi potongan teks yang lebih terstruktur dan terkelola. Variasi panjang potongan ini dirancang untuk mempermudah proses ekstraksi fitur *stylometry* serta memberikan ruang evaluasi terhadap kinerja model klasifikasi dan deteksi plagiarisme. Dengan adanya perbandingan pada beberapa skala *chunk*, penelitian dapat menguji sejauh mana panjang teks mempengaruhi keakuratan identifikasi penulis maupun efektivitas pengukuran tingkat kemiripan antar teks.

Struktur dataset yang sudah di *chunk* dan disimpan dalam format CSV, dengan struktur tabel sebagai berikut:

Tabel 3. 1 Contoh Struktur Tabel

Id	Author	Novel	Id_Chunks	Chunks
A1	Nama Penulis	Judul Novel	A1N1-1	Isi text per chunk
A1	…	…	A1N1-2	…
A2	…	…	…	…

Tabel 3.1 menampilkan potongan-potongan novel yang telah dibersihkan dari elemen yang tidak relevan seperti tanda baca, angka, maupun karakter khusus sehingga hanya menyisakan teks murni. Setiap baris pada tabel merepresentasikan satu segmen teks hasil *chunking* dengan panjang tertentu (misalnya 1000 atau 5000 kata) yang telah melalui tahap normalisasi, seperti konversi huruf menjadi bentuk kecil (*lowercasing*) dan penghapusan spasi berlebih. Kolom id dan id_chunks digunakan sebagai identitas unik setiap segmen, sedangkan kolom *author* dan novel menunjukkan nama penulis serta judul karya asalnya. Kolom *cleaned*_text berisi teks novel yang sudah diproses sehingga siap digunakan dalam tahap analisis lebih lanjut, baik untuk ekstraksi fitur *stylometry* maupun perhitungan kesamaan semantik.

3.3.3 Ekstraksi Fitur Stylometry

Pada tahap ini dilakukan ekstraksi fitur linguistik dari setiap *chunk* teks. Fitur-fitur yang digunakan antara lain rata-rata panjang kalimat, rata-rata panjang kata,

type-token ratio, hapax legomena ratio, distribusi kata fungsi, *entropi n-gram* karakter, hingga informasi *part-of-speech* (POS Tag). Fitur-fitur ini dianggap sebagai sidik jari penulis (gaya khas tulisan) yang relatif konsisten meskipun topik tulisan berubah. Hasil ekstraksi fitur *stylometry* yang disimpan dalam bentuk vektor numerik selanjutnya digunakan dalam model klasifikasi berbasis *Support Vector Machine* (SVM) untuk memprediksi penulis teks.

3.3.4 Analisis Semantic Similarity (TF-IDF & SBERT)

Analisis kesamaan semantik dilakukan dengan dua pendekatan utama untuk mengevaluasi sejauh mana teks tiruan mempertahankan makna dari teks asli.

Pendekatan pertama, TF-IDF (*Term Frequency–Inverse Document Frequency*), digunakan sebagai baseline. TF-IDF merepresentasikan teks dalam bentuk vektor numerik berdasarkan frekuensi kata dan seberapa unik kata tersebut di seluruh dokumen. Dengan demikian, kata-kata yang sering muncul di seluruh korpus akan memiliki bobot rendah, sementara kata-kata yang lebih khas terhadap sebuah dokumen memiliki bobot lebih tinggi. Setelah representasi vektor diperoleh, kemiripan antar teks dihitung menggunakan cosine similarity, yang mengukur sejauh mana arah vektor teks satu sejajar dengan teks lainnya. Pendekatan ini efektif untuk menangkap kesamaan berbasis kata, tetapi memiliki keterbatasan dalam memahami konteks atau makna kalimat secara mendalam, terutama jika teks mengalami parafrasa atau perubahan struktur kalimat.

Pendekatan kedua, SBERT (*Sentence*-BERT), digunakan untuk menghasilkan representasi semantik yang lebih kaya. SBERT merupakan pengembangan dari BERT (*Bidirectional Encoder Representations from Transformers*) yang mampu menghasilkan *embedding* untuk tingkat kalimat atau paragraf, sehingga dapat menangkap konteks dan makna kata dalam hubungannya dengan kata lain di dalam kalimat. Representasi ini memungkinkan perhitungan *cosine similarity* antara teks asli dan tiruan untuk mendeteksi kesamaan makna yang lebih halus, termasuk pada kasus parafrasa, sinonim, atau perubahan susunan kalimat. Dengan menggunakan SBERT, sistem tidak hanya mengandalkan kemiripan kata secara literal, tetapi juga mempertimbangkan konteks dan hubungan semantik antar kata.

3.3.5 Integrasi Stylometry & Semantic Similarity

Sistem ini mengintegrasikan hasil analisis dari dua pendekatan, yaitu *stylometry* dan *semantic similarity*, untuk memperoleh deteksi yang lebih komprehensif terhadap plagiarisme dan identifikasi penulis. Dari sisi *stylometry*, sistem menghasilkan prediksi penulis teks berdasarkan pola linguistik, seperti panjang kalimat, pemilihan kata, tanda baca, dan struktur sintaksis, menggunakan model SVM (Support Vector Machine). Prediksi ini menunjukkan sejauh mana gaya penulisan suatu teks sesuai dengan ciri khas seorang penulis.

Dari sisi *semantic similarity*, sistem menghitung skor kesamaan makna antar teks, menggunakan kombinasi TF-IDF dan SBERT. Skor ini memberikan informasi mengenai sejauh mana isi teks tiruan mempertahankan makna atau ide dari teks asli, termasuk ketika terjadi parafrasa atau variasi redaksi kalimat. Dengan demikian, *semantic similarity* melengkapi analisis *stylometry* yang fokus pada gaya penulisan, sehingga setiap kesamaan konten dapat terdeteksi meskipun gaya penulisannya berbeda.

Integrasi kedua pendekatan ini menghasilkan dua keluaran utama:

1. Prediksi penulis berdasarkan model SVM, yang menunjukkan kecocokan teks dengan gaya penulis asli.
2. Skor plagiarisme, yang menggabungkan informasi dari *stylometry* dan *semantic similarity* untuk menilai tingkat kesamaan antara teks asli dan teks tiruan secara menyeluruh.

Pendekatan terpadu ini memastikan bahwa sistem tidak hanya mampu mengidentifikasi gaya penulisan, tetapi juga menilai kesamaan makna secara mendalam. Hal ini meningkatkan akurasi deteksi plagiarisme, terutama pada kasus yang kompleks, seperti parafrasa atau teks yang meniru gaya penulis lain namun memiliki makna berbeda. Dengan demikian, integrasi *stylometry* dan *semantic similarity* menjadikan sistem lebih adaptif, robust, dan komprehensif dalam mendukung penelitian *authorship attribution* maupun deteksi plagiarisme modern.

3.4 Evaluasi Kinerja Model Klasifikasi

Evaluasi kinerja model klasifikasi dilakukan untuk mengukur sejauh mana sistem mampu mengenali penulis teks dengan benar berdasarkan fitur *stylometry*

yang telah diekstraksi. Pada penelitian ini, model klasifikasi yang digunakan adalah *Support Vector Machine* (SVM) karena kemampuannya dalam menangani data berdimensi tinggi serta menghasilkan *hyperplane* optimal untuk memisahkan kelas.

3.4.1 Algoritma Support Vector Machine (SVM)

SVM adalah algoritma *supervised learning* yang bekerja dengan mencari *hyperplane* terbaik untuk memisahkan data ke dalam kelas berbeda. Dalam konteks penelitian ini, SVM digunakan untuk melakukan *authorship attribution*, yaitu prediksi penulis teks berdasarkan fitur *stylometry*. Setiap *chunk* teks direpresentasikan sebagai vektor fitur, seperti panjang kalimat, panjang kata, type-token ratio, hapax legomena ratio, distribusi POS tag, dan fitur lain yang relevan. Model SVM kemudian membangun *hyperplane* untuk memaksimalkan margin antar kelas, sehingga teks dari penulis berbeda dapat dipisahkan secara optimal.

Secara matematis, SVM membentuk sebuah fungsi keputusan berupa *hyperplane*:

$$f(x) = w \,.\, x + b$$

dengan w adalah vector bobot, x vector fitur, dan b Adalah bias.

Hyperplane tersebut digunakan untuk memisahkan data ke dalam kelas yang berbeda dengan memaksimalkan margin (jarak antara data terdekat dengan garis pemisah).

Pemilihan SVM didasarkan pada beberapa alasan yaitu kemampuannya menangani data berdimensi besar, stabilitas dalam generalisasi, serta performa baik pada klasifikasi dengan margin jelas. *Kernel* RBF (*Radial Basis Function*) digunakan karena distribusi fitur penulis lebih mudah dipisahkan dalam ruang non-linear berdimensi tinggi dibandingkan kernel linear atau polynomial.

3.4.2 Persiapan Data dan Standardisasi

Data dibagi menjadi set latih dan uji menggunakan `train_test_split` dengan parameter `stratify=y` agar distribusi label penulis seimbang. Semua fitur kemudian distandarisasi menggunakan *StandardScaler*, sehingga setiap fitur memiliki mean 0 dan standar deviasi 1. Standardisasi ini penting karena SVM, khususnya dengan kernel RBF, sensitif terhadap skala fitur dan dapat memengaruhi posisi *hyperplane* optimal.

3.4.3 Hyperparameter Tuning

Untuk meningkatkan kinerja model, SVM dilatih menggunakan *GridSearchCV* dengan *StratifiedKFold 10fold* untuk menemukan kombinasi *hyperparameter* terbaik, yaitu `C` (penalti kesalahan), `gamma` (skala kernel RBF), `kernel` (linear, rbf, poly), dan `class_weight` (*None* atau *balanced*). Pendekatan ini memungkinkan pencarian parameter secara sistematis, sehingga model dapat menyesuaikan *margin* pemisahan antar kelas dengan optimal dan mengurangi risiko *overfitting*, sekaligus memastikan akurasi tinggi pada data uji yang distribusi labelnya seimbang.

3.4.4 Metrik Evaluasi (Akurasi, Presisi, Recall, F1-Score)

Kinerja model klasifikasi dievaluasi menggunakan empat metrik utama:

1. Akurasi

   Mengukur proporsi prediksi benar dibandingkan dengan seluruh data uji. Akurasi menunjukkan seberapa tepat model dalam mengklasifikasikan teks ke penulis yang sesuai.

2. Presisi

   Menunjukkan ketepatan prediksi untuk setiap kelas, yaitu seberapa banyak teks yang benar-benar ditulis oleh seorang penulis dari seluruh teks yang diprediksi sebagai miliknya.

3. *Recall*

   Mengukur kemampuan model dalam menemukan semua teks yang benar-benar ditulis oleh seorang penulis. Nilai recall tinggi berarti model jarang melewatkan teks yang seharusnya masuk ke kelas tertentu.

4. F1-*Score*

   Merupakan *harmonic mean* dari presisi dan *recall*, sehingga memberikan gambaran seimbang antara ketepatan dan kelengkapan model.

Dengan penggunaan metrik-metrik tersebut, kinerja SVM dalam melakukan atribusi kepenulisan dapat dinilai secara menyeluruh, tidak hanya dari sisi ketepatan prediksi, tetapi juga dari sisi keseimbangan antara presisi dan recall pada masing-masing penulis.

3.4.5 Penyimpanan Model

Model SVM terbaik beserta scaler disimpan menggunakan joblib.dump, sehingga dapat digunakan kembali tanpa perlu melatih ulang. Hasil *GridSearchCV* juga disimpan dalam *joblib.dump* untuk keperluan dokumentasi dan analisis lebih lanjut.

Dengan pendekatan ini, evaluasi kinerja model tidak hanya mengukur akurasi global, tetapi juga menilai performa per penulis dan kemampuan generalisasi model terhadap teks baru. Kombinasi metrik evaluasi dan visualisasi mendukung analisis menyeluruh, serta memberikan landasan yang kuat untuk integrasi lebih lanjut dengan metode semantic similarity dalam deteksi plagiarisme.

3.5 Evaluasi Kinerja Deteksi Plagiarisme

Evaluasi kinerja deteksi plagiarisme dilakukan untuk mengukur sejauh mana sistem mampu mendeteksi teks tiruan yang meniru gaya dan makna teks asli. Penilaian dilakukan menggunakan dua pendekatan utama, yaitu *stylometry* dan *semantic similarity*, yang dianalisis baik secara terpisah maupun secara kombinasi. Selain itu, analisis dilakukan pada berbagai ukuran *chunk* teks (1000, 5000, dan 10000 kata) untuk menilai pengaruh panjang potongan teks terhadap akurasi deteksi.

3.5.1 Nilai Stylometry

Nilai *stylometry* diperoleh dari perhitungan kesamaan vektor fitur gaya penulisan antar chunk teks. Fitur yang dianalisis meliputi panjang kalimat rata-rata, panjang kata rata-rata, *type-token ratio*, *hapax legomena ratio*, *vowel ratio, punctuation ratio*, *stopword ratio*, *function word ratio*, *char n-gram entropy*, panjang *chunk* (jumlah kata).

Vektor fitur setiap *chunk* teks tiruan dibandingkan dengan dataset teks asli menggunakan *cosine similarity*, setelah distandarisasi dengan *StandardScaler*. Skor maksimal per chunk digunakan untuk mewakili tingkat kemiripan dengan teks asli, kemudian dirata-rata untuk memperoleh skor keseluruhan (*avg_stylo_score*). Semakin tinggi nilai ini, semakin besar kemungkinan teks tersebut meniru gaya penulis asli.

3.5.2 Nilai Semantic Similarity

Semantic similarity dihitung menggunakan *Sentence*-BERT (SBERT) sebagai representasi teks berbasis *embedding* yang mempertimbangkan konteks antar kalimat. Teks dibagi menjadi *chunk* 1000 kata, dan setiap *chunk* dikonversi menjadi *embedding* vektor. Kemiripan antar teks tiruan dan teks asli dihitung menggunakan *cosine similarity*, dengan skor maksimal antar *chunk* sebagai representasi kemiripan. Rata-rata skor maksimal per *chunk* dijadikan *avg_sem_score*, yang menunjukkan kedekatan makna teks tiruan dengan teks asli. Dengan pendekatan ini, sistem dapat mendeteksi plagiarisme berbentuk parafrasa atau perubahan struktur kata, karena kemiripan makna tetap diperhitungkan.

3.5.3 Analisis Kombinasi Stylometry & Semantic

Analisis kombinasi dilakukan dengan mengintegrasikan skor *stylometry* dan skor *semantic similarity* untuk menghasilkan deteksi plagiarisme yang lebih komprehensif. *Stylometry* berperan dalam memastikan kesamaan gaya tulis yang relatif konsisten dari seorang penulis, sedangkan *semantic similarity* berfungsi dalam mendeteksi kesamaan makna antar teks. Dengan menggabungkan keduanya, sistem tidak hanya mampu mengidentifikasi plagiarisme berbasis kata demi kata, tetapi juga plagiarisme yang lebih kompleks seperti parafrasa maupun modifikasi struktur kalimat.

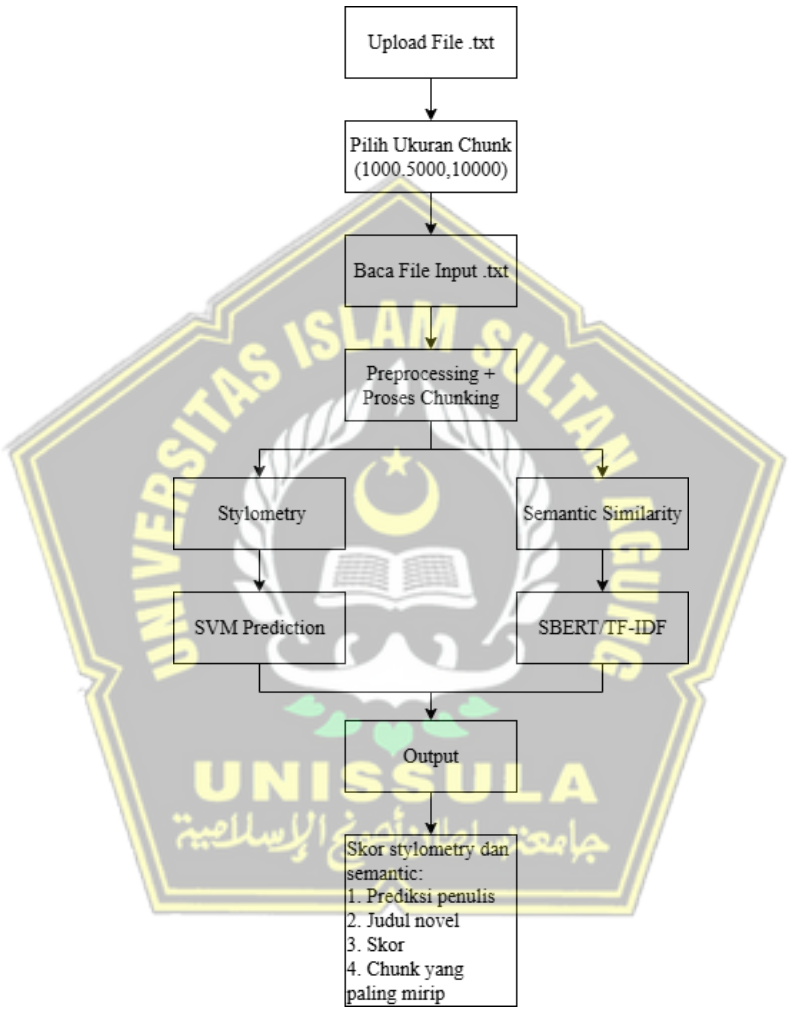
Hasil analisis disimpan dalam file .csv, yang memuat kolom *chunk_size*, *gpt_author*, *gpt_novel*, *avg_stylo_score*, dan *avg_sem_score*. Pendekatan berbasis chunk memungkinkan evaluasi baik untuk teks pendek maupun panjang, serta analisis per penulis dan per novel.

Dengan metode ini, sistem tidak hanya mendeteksi plagiarisme berbasis kata demi kata, tetapi juga plagiarisme yang lebih kompleks, seperti parafrasa, modifikasi kalimat, dan penyesuaian kosakata. Analisis kombinasi ini meningkatkan keakuratan dan ketelitian deteksi plagiarisme secara signifikan dibandingkan menggunakan salah satu metode saja.

3.6 Alur Sistem Antarmuka

Alur sistem antarmuka aplikasi ditunjukkan pada Gambar 3.4. Diagram tersebut menggambarkan langkah-langkah interaksi pengguna dengan sistem dalam

proses identifikasi penulis dan deteksi plagiarisme. Adapun penjelasan setiap tahap adalah sebagai berikut:



Gambar 3. 4 Alur system antar muka

1. Upload File .txt

   Pengguna mengunggah file teks berformat .txt yang akan dianalisis. File ini berisi naskah novel atau teks uji yang menjadi objek penelitian.

2. Pilih Ukuran Chunk (1000, 5000, 10000 kata)

Pengguna memilih ukuran chunk atau potongan teks untuk analisis. Pilihan ukuran ini memungkinkan sistem melakukan segmentasi teks agar perhitungan fitur *stylometry* maupun *semantic similarity* lebih terstruktur.

3. Baca File Input .txt

   Sistem membaca isi file yang telah diunggah untuk memastikan data dapat diproses lebih lanjut.

4. *Preprocessing* dan Proses *Chunking*

   File teks dibersihkan melalui *preprocessing* (pembersihan tanda baca, normalisasi huruf, tokenisasi kata), kemudian dipotong menjadi beberapa *chunk* sesuai ukuran yang dipilih.

5. *Stylometry*

   Hasil *chunking* dianalisis menggunakan pendekatan *stylometry*. Sistem mengekstraksi fitur-fitur *linguistik* (panjang kalimat, panjang kata, frekuensi tanda baca, dll.) untuk mengidentifikasi gaya penulisan.

6. *Semantic Similarity* (TF-IDF / SBERT)

   Sistem merepresentasikan teks dalam bentuk vektor semantik menggunakan TF-IDF dan SBERT. Kemudian, dihitung skor kesamaan antar teks menggunakan *cosine similarity*.

7. SVM *Prediction*

   Model SVM dilatih menggunakan fitur *stylometry* untuk memprediksi penulis dari teks yang diuji.

8. *Output*

   Sistem menampilkan hasil akhir berupa skor *stylometry* dan *semantic* similarity yang mencakup prediksi penulis teks, judul novel asal, skor kesamaan teks dan *chunk* yang memiliki kemiripan paling tinggi.

Dengan alur ini, antarmuka aplikasi tidak hanya mempermudah pengguna dalam menjalankan analisis, tetapi juga menyajikan hasil akhir secara jelas dan informatif, sehingga dapat mendukung proses identifikasi penulis sekaligus deteksi plagiarisme.

BAB IV
HASIL DAN ANALISIS PENELITIAN

4.1 Hasil

Mengacu pada gambar 3.1 alur perancangan sistem, penelitian ini melalui beberapa tahapan untuk menghasilkan model klasifikasi penulis dan deteksi plagiarisme. Proses tersebut mencakup pengolahan teks novel berbahasa Inggris, *chunking* teks, ekstraksi fitur *stylometry*, pelatihan model *Support Vector Machine* (SVM), serta perhitungan kesamaan semantik menggunakan SBERT.

4.2 Deskripsi Data Penelitian

4.2.1 Distribusi Data per Penulis

Data penelitian berdasarkan jumlah novel, hasil proses *chunking* dengan tiga ukuran berbeda (1000 kata, 5000 kata, dan 10000 kata), serta total proporsi chunk yang dihasilkan dari masing-masing penulis.

Tabel 4. 1 Tabel Distribusi per penulis

Penulis	1000 Kata	5000 Kata	10000 Kata
Jane Austen	338	69	35
Mark Twain	243	50	26
Bram Stoker	154	33	18
Herbert George Wells	112	24	12
Marry Shelley	112	23	12
Total	**959**	**199**	**103**

Dari Tabel 4.1 hasil proses *chunking* novel, diperoleh jumlah potongan teks (*chunk*) yang bervariasi untuk setiap penulis dan ukuran *chunk*. Pada chunk 1000 kata, penulis dengan jumlah data terbanyak adalah Jane Austen dengan 338 *chunk*, diikuti oleh Mark Twain sebanyak 243 *chunk*, Bram Stoker sebanyak 154 *chunk*, serta Herbert George Wells dan Mary Shelley masing-masing 112 *chunk*. Sementara itu, pada *chunk* 5000 kata, jumlah data yang diperoleh mengalami pengurangan sesuai ukuran potongan teks. Jane Austen tetap mendominasi dengan 69 chunk, disusul Mark Twain sebanyak 50 chunk, Bram Stoker 33 chunk, Herbert George Wells 24 chunk, dan Mary Shelley 23 chunk. Pada chunk 10000 kata, distribusi data

semakin kecil dengan Jane Austen menghasilkan 35 chunk, Mark Twain 26 chunk, Bram Stoker 18 *chunk*, serta Herbert George Wells dan Mary Shelley masing-masing 12 chunk.

Distribusi ini menunjukkan bahwa jumlah *chunk* sangat dipengaruhi oleh panjang novel asli dari masing-masing penulis. Penulis dengan karya yang lebih panjang secara otomatis menghasilkan *chunk* yang lebih banyak ketika dilakukan pemotongan. Hal ini penting karena distribusi data yang seimbang antar penulis akan berpengaruh pada keakuratan model *machine learning* yang digunakan pada tahap berikutnya.

4.2.2 Pembagian Data (Latih, Uji, Validasi)

Setelah diperoleh jumlah *chunk* dari masing-masing penulis, data kemudian dibagi menjadi tiga bagian, yaitu data latih, data validasi, dan data uji. Pada dataset 1000 kata dengan total 959 *chunk*, data dibagi menjadi 767 untuk latih, 96 untuk validasi, dan 96 untuk uji. Pada dataset 5000 kata dengan total 199 *chunk*, diperoleh 159 data latih, 20 data validasi, dan 20 data uji. Sementara pada dataset 10000 kata dengan total 103 chunk, diperoleh 82 data latih, 10 data validasi, dan 11 data uji.

Tabel 4. 2 Tabel pembagian data

Dataset	Total Data	Latih(80%)	Validasi(10%)	Uji (10%)
1000 Kata	959	767	96	96
5000 Kata	199	159	20	20
10000 Kata	103	82	10	11

Pembagian ini mengikuti skema 80% untuk data latih, 10% untuk validasi, dan 10% untuk uji, dengan mempertahankan stratifikasi berdasarkan penulis agar distribusi data tetap seimbang. Dengan strategi ini, data latih digunakan untuk membangun model, data validasi digunakan untuk menyesuaikan parameter dan mencegah *overfitting*, sedangkan data uji digunakan untuk mengukur performa akhir model secara objektif.

4.3 Hasil Implementasi Sistem

4.3.1 Hasil Preprocessing Teks

Tahap *preprocessing* dilakukan untuk membersihkan teks novel dari berbagai elemen yang tidak relevan, seperti nomor halaman, catatan *Project Gutenberg*, serta karakter khusus yang tidak berhubungan dengan isi narasi. Proses ini meliputi *case* folding (mengubah seluruh huruf menjadi huruf kecil), penghapusan tanda baca dan angka, serta normalisasi spasi. Dengan *preprocessing*, teks menjadi lebih konsisten dan siap untuk dilakukan analisis lebih lanjut. Hasil *preprocessing* menunjukkan bahwa teks novel dari masing-masing penulis telah seragam dalam format, sehingga memudahkan dalam tahap ekstraksi fitur *stylometry* maupun analisis kesamaan semantik.

4.3.2 Hasil Chunking Dokumen

Novel yang telah dipreproses kemudian dibagi menjadi potongan (*chunk*) dengan ukuran berbeda, yaitu 1000 kata, 5000 kata, dan 10000 kata. Hasil *chunking* menghasilkan variasi jumlah potongan pada tiap penulis, bergantung pada panjang novel yang dimiliki. Misalnya, Jane Austen memiliki jumlah *chunk* terbanyak dibandingkan penulis lain karena novel yang digunakan relatif panjang. Proses *chunking* ini bertujuan untuk menyediakan unit analisis yang lebih kecil dan seragam, sehingga model dapat membandingkan gaya penulisan secara lebih efektif. Hasil distribusi chunk per penulis telah ditunjukkan pada Tabel 4.1 Tabel Distribusi per penulis dan Tabel 4.2 Tabel pembagian data.

4.3.3 Hasil Ekstraksi Fitur Stylometry

Proses ekstraksi fitur *stylometry* dilakukan untuk memperoleh ciri khas gaya penulisan setiap penulis dari potongan teks hasil chunking. Fitur-fitur yang diekstraksi meliputi frekuensi kata umum (*function words),* panjang rata-rata kalimat, panjang rata-rata kata, distribusi tanda baca, serta proporsi huruf kapital. Fitur ini dipilih karena bersifat independen dari topik teks, sehingga lebih mencerminkan gaya menulis seorang penulis dibandingkan isi ceritanya. Hasil ekstraksi fitur disimpan ke dalam tiga berkas CSV, yaitu 1000_features.csv, 5000_features.csv, dan 10000_features.csv, yang masing-masing merepresentasikan potongan teks dengan panjang 1000 kata, 5000 kata, dan 10000

kata. Setiap baris pada dataset berisi informasi *id_chunks*, *author*, *novel*, serta nilai fitur *stylometry* yang telah dihitung.

Hasil ekstraksi fitur *stylometry* menghasilkan sejumlah metrik linguistik yang merepresentasikan gaya penulisan setiap penulis. Fitur yang dianalisis meliputi rata-rata panjang kalimat (*average sentence length*), panjang kata (*average word length)*, variasi kosakata melalui *type token ratio* (TTR) dan *hapax legomena ratio*, proporsi huruf vokal (*vowel ratio*), tanda baca (*punctuation ratio*), kata umum (*stopword ratio*), kata fungsi (*function word ratio*), serta kompleksitas pola karakter yang diukur dengan *char n-gram entropy*. Proses ekstraksi dilakukan dengan membagi teks ke dalam potongan (*chunk*) berukuran 1000, 5000, dan 10000 kata, sehingga menghasilkan 959 *chunk* untuk 1000 kata, 199 *chunk* untuk 5000 kata, dan 103 *chunk* untuk 10000 kata.

Pada *chunk* berukuran 1000 kata, nilai *type token ratio* relatif tinggi dengan rata-rata 0.48 dan dapat mencapai maksimum 0.87. Hal ini menunjukkan bahwa potongan teks pendek memiliki variasi kosakata yang lebih besar, karena setiap kata baru memberi kontribusi signifikan terhadap rasio TTR. *Hapax legomena ratio* pada ukuran ini juga tinggi, dengan rata-rata 0.32 dan maksimum hingga 0.77, menandakan banyak kata unik yang hanya muncul sekali dalam potongan teks kecil. Rata-rata panjang kata berada pada kisaran 4.6 huruf per kata, sedangkan panjang kalimat rata-rata mendekati 1000 kata sesuai pembagian chunk.

Berbeda dengan *chunk* berukuran 5000 kata, nilai TTR turun signifikan menjadi rata-rata sekitar 0.30 dengan hapax rata-rata 0.18. Penurunan ini dapat dijelaskan karena semakin panjang teks, semakin besar kemungkinan kata-kata yang sama muncul berulang, sehingga menurunkan proporsi kosakata unik. Meskipun begitu, beberapa fitur lain seperti stopword ratio (rerata 0.42) dan *function word ratio* (sekitar 0.098) tetap relatif stabil. Hal ini menunjukkan bahwa penggunaan kata hubung dan kata fungsi cenderung konsisten dalam setiap potongan teks, sehingga bisa dianggap sebagai ciri khas yang stabil dari penulis.

Pada *chunk* 10000 kata, pola yang sama semakin terlihat jelas. Nilai TTR rata-rata hanya sekitar 0.23, sedangkan *hapax legomena* semakin kecil dengan nilai rata-rata 0.13. Penurunan ini memperlihatkan bahwa semakin panjang teks, semakin

dominan repetisi kata yang digunakan. Sebaliknya, nilai *char n-gram entropy* justru meningkat, dengan rata-rata sekitar 11.04. Hal ini menandakan bahwa semakin panjang teks, distribusi pola karakter semakin kompleks dan lebih representatif terhadap keseluruhan gaya penulisan. Rata-rata panjang kata tetap berada pada kisaran 4.4 huruf, memperlihatkan bahwa faktor ini relatif stabil dan tidak banyak dipengaruhi oleh ukuran chunk.

Secara keseluruhan, hasil ekstraksi fitur *stylometry* menunjukkan bahwa ukuran *chunk* berpengaruh terhadap nilai-nilai metrik tertentu. Fitur yang terkait dengan variasi kosakata (TTR dan *hapax legomena*) cenderung lebih tinggi pada potongan kecil dan semakin menurun pada potongan besar. Sebaliknya, fitur yang bersifat struktural dan konsisten seperti penggunaan stopword, *function word*, serta rata-rata panjang kata tetap stabil pada semua ukuran chunk. Hal ini menguatkan asumsi bahwa ciri khas gaya penulis dapat dilihat dari pola penggunaan kata fungsi dan *stopword*, sementara variasi kosakata lebih dipengaruhi oleh panjang teks yang dianalisis.

Tabel 4. 3 Tabel Ringkasan Dataset Fitur *Stylometry*

Dataset Fitur	**Jumlah Chunk**	**Jumlah Fitur**	**Keterangan Utama**
1000 features	959	14	Detail tinggi, cocok untuk analisis granular gaya penulis
5000 features	199	14	Representasi lebih stabil dan mengurangi noise dari variasi lokal
10000 features	103	14	Memberikan ciri umum gaya penulis, namun data lebih terbatas

Tabel 4.3 merangkum tiga dataset hasil ekstraksi fitur *stylometry* yang dibedakan berdasarkan ukuran potongan teks (*chunk size*). Setiap dataset terdiri dari 14 fitur utama yang menggambarkan aspek linguistik penulis, seperti panjang kalimat, panjang kata, variasi kosakata, penggunaan *stopword* dan kata fungsi, hingga kompleksitas pola karakter (*char n-gram entropy*). Perbedaan utama antar dataset terletak pada jumlah *chunk* yang dihasilkan dari proses pemotongan teks. Dataset 1000 kata memberikan detail tinggi, dataset 5000 kata menawarkan keseimbangan, sedangkan dataset 10000 kata menampilkan gambaran umum gaya penulis.

Hasil ekstraksi fitur *stylometry* ditunjukkan dengan skor rata-rata *stylometry* untuk tiap penulis tiruan.

Tabel 4. 4 Rata-rata Skor Stylometry per Penulis Tiruan

Penulis Tiruan	1000 kata	5000 kata	10000 kata
Jane Austen	0.874	0.815	0.754
Mark Twain	0.874	0.773	0.676
Bram Stoker	0.891	0.781	0.725
Herbert George Wells	0.918	0.840	0.787
Mary Shelley	0.860	0.758	0.679

Dari Tabel 4.4 terlihat bahwa nilai rata-rata stylometry bervariasi antar penulis tiruan. Herbert George Wells (tiruan) memiliki skor tertinggi (0.918 pada *chunk* 1000 kata dan tetap tinggi pada ukuran lebih besar), menunjukkan bahwa gaya penulisan GPT untuk penulis ini lebih konsisten. Sementara itu, Mary Shelley (tiruan) memperoleh skor paling rendah (0.679 pada *chunk* 10000 kata), menandakan bahwa teks tiruan untuk penulis ini kurang stabil dalam mempertahankan ciri khas gaya penulisan. Secara umum, semakin besar ukuran chunk, skor *stylometry* cenderung menurun, yang mengindikasikan semakin banyak variasi gaya yang muncul dalam potongan teks panjang.

4.3.4 Hasil Analisis Semantic Similarity (TF-IDF dan SBERT)

Analisis *semantic similarity* dilakukan menggunakan TF-IDF dan SBERT. Nilai berikut merupakan skor rata-rata *semantic similarity*.

Tabel 4. 5 Rata-rata Skor Semantic Similarity per Penulis Tiruan

Penulis Tiruan	1000 kata	5000 kata	10000 kata
Jane Austen	0.639	0.624	0.617
Mark Twain	0.573	0.528	0.517
Bram Stiker	0.576	0.529	0.514
Herbert George Wells	0.590	0.539	0.520
Mary Shelley	0.561	0.512	0.496

Tabel 4.5 menampilkan skor *semantic similarity* rata-rata untuk teks tiruan. Secara umum, skor semantic similarity berkisar antara 0.49 hingga 0.63, dengan tren penurunan ketika ukuran *chunk* semakin besar. Jane Austen (tiruan) menunjukkan skor *semantic similarity* tertinggi (0.639 pada chunk 1000 kata), menandakan konsistensi semantik relatif lebih baik. Sebaliknya, Mary Shelley (tiruan) menempati posisi terendah (0.496 pada chunk 10000 kata), sehingga dapat dikatakan teks tiruan untuk Mary Shelley lebih sulit dipertahankan konsistensi maknanya.

Berdasarkan hasil analisis *semantic similarity* pada teks tiruan dengan ukuran *chunk* berbeda (1000, 5000, dan 10000 kata), terlihat adanya variasi skor kesamaan semantik antarpenulis dan antarukuran *chunk*. Pada *chunk* berukuran 1000 kata, skor rata-rata semantic similarity berada pada rentang 0.51–0.64, dengan nilai tertinggi ditunjukkan oleh karya tiruan Mary Shelley berjudul *Eleanor, or the Ruins of Silence* (0.6483) dan karya tiruan Jane Austen seperti *The Sisters of Ashbourne* (0.6474). Hal ini mengindikasikan bahwa pada potongan teks yang lebih kecil, model GPT mampu menghasilkan representasi semantik yang cukup dekat dengan teks asli.

Namun, ketika ukuran chunk diperbesar menjadi 5000 dan 10000 kata, skor kesamaan semantik cenderung menurun, terutama pada beberapa penulis seperti Mary Shelley (*The Towers of Bellacqua hanya* 0.3878 pada *chunk* 5000 dan 0.3871 pada *chunk* 10000). Penurunan skor ini menunjukkan bahwa semakin panjang teks, semakin besar pula peluang terjadinya variasi konten dan struktur naratif yang menyebabkan perbedaan makna lebih terlihat dibandingkan teks asli. Secara umum, skor *semantic similarity* lebih stabil pada Jane Austen dan Herbert G. Wells, sementara Mark Twain dan Mary Shelley lebih fluktuatif.

4.3.5 Integrasi Stylometry & Semantic Similarity

Integrasi dilakukan dengan menggabungkan skor *stylometry* dan *semantic similarity* untuk mengevaluasi konsistensi penulis tiruan.

Tabel 4. 6 Ringkasan Integrasi Stylometry & Semantic Similarity

Penulis Tiruan	Rata-rata Stylometry	Rata-rata Semantic
Jane Austen	0.814	0.627
Mark Twain	0.774	0.539
Bram Stoker	0.799	0.540
Herbert George Wells	0.848	0.550
Mary Shelley	0.766	0.523

Integrasi antara *stylometry* dan *semantic similarity* menghasilkan gambaran lebih jelas mengenai kualitas teks tiruan. Berdasarkan Tabel 4.6, Herbert George Wells (tiruan) menonjol dengan kombinasi skor stylometry (0.848) dan *semantic similarity* (0.550), sehingga bisa dikategorikan sebagai penulis tiruan yang paling konsisten. Sebaliknya, Mary Shelley (tiruan) kembali menempati posisi terbawah dengan skor *stylometry* (0.766) dan *semantic similarity* (0.523). Hal ini memperlihatkan bahwa GPT lebih mudah meniru gaya penulisan Wells dibandingkan Shelley, baik dari sisi gaya bahasa maupun kesesuaian semantik.

4.4 Hasil Evaluasi Model Klasifikasi

4.4.1 Hasil Pelatihan dengan Support Vector Machine (SVM)

Pelatihan model klasifikasi menggunakan *Support Vector Machine* (SVM) dilakukan dengan tiga variasi ukuran *chunk* teks, yaitu 1000 kata, 5000 kata, dan 10000 kata. Parameter model ditentukan melalui *grid search*, dan hasil terbaik diperoleh dengan penggunaan kernel RBF (Radial Basis Function) pada semua percobaan. Nilai parameter C bervariasi antara 10 hingga 100, sementara nilai gamma tetap menggunakan pengaturan scale. Pemilihan kernel RBF menunjukkan bahwa distribusi data penulis tiruan lebih mudah dipisahkan dalam ruang *non-linear* berdimensi tinggi, dibandingkan dengan *kernel linear* atau *polynomial*. Hal ini sejalan dengan karakteristik data *stylometry* dan semantic *similarity* yang kompleks.

4.4.2 Evaluasi Kinerja Model (Akurasi, Presisi, Recall, F1-Score)

Setelah proses pelatihan dengan algoritma *Support Vector Machine* (SVM), tahap berikutnya adalah mengevaluasi kinerja model menggunakan beberapa metrik utama, yaitu akurasi, presisi, *recall, dan F1-score*. Evaluasi ini dilakukan pada tiga variasi ukuran *chunk* teks, yaitu 1000 kata, 5000 kata, dan 10000 kata. Pemilihan variasi ukuran *chunk* bertujuan untuk mengetahui sejauh mana jumlah kata dalam setiap segmen teks memengaruhi kinerja model dalam mengenali gaya penulisan masing-masing penulis.

Berikut ini dipaparkan hasil evaluasi model pada setiap ukuran *chunk*.

1. Hasil Evaluasi Model dengan *Chunk* 1000 Kata

Tabel 4. 7 Evaluasi Model dengan *Chunk* 1000 Kata

Penulis	Precision	Recall	F1-Score	Support
Bram Stoker	0.7647	0.8387	0.8000	31
Herbert George Wells	0.8500	0.7727	0.8095	22
Jane Austen	0.9286	0.9559	0.9420	68
Mark Twain	0.7679	0.8776	0.8190	49
Mary Shelley	0.9167	0.5000	0.6471	22
Accuracy			**0.8438**	192

Pada dataset *chunk* 1000 kata dengan total 959 data (767 latih, 96 validasi, 96 uji), model SVM menghasilkan akurasi 84.38%. Hasil ini menunjukkan bahwa model mampu membedakan gaya penulisan dengan cukup baik pada potongan teks pendek. Jika dilihat lebih detail, Jane Austen memiliki kinerja terbaik dengan *precision* 0.92, *recall* 0.95, dan f1-*score* 0.94, menandakan gaya penulisannya sangat konsisten dan mudah dikenali. Mark Twain juga cukup stabil (f1-*score* 0.81), sedangkan Bram Stoker dan Herbert George Wells berada pada kategori menengah (f1-*score* 0.80 dan 0.81). Namun, kelemahan paling terlihat ada pada Mary Shelley dengan *recall* hanya 0.50, menunjukkan bahwa model sering salah mengenali teks miliknya sebagai penulis lain.

2. Hasil Evaluasi Model dengan *Chunk* 5000 Kata

Tabel 4. 8 Evaluasi Model dengan Chunk 5000 Kata

Penulis	Precision	Recall	F1-Score	Support
Bram Stoker	0.7143	0.7143	0.7143	7
Herbert George Wells	0.5714	0.8000	0.6667	5
Jane Austen	0.9333	1.0000	0.9655	14
Mark Twain	0.8889	0.8000	0.8421	10
Mary Shelley	1.0000	0.5000	0.6667	4
Accuracy			**0.8250**	40

Pada dataset chunk 5000 kata dengan total 199 data (159 latih, 20 validasi, 20 uji), akurasi model berada di angka 82.5%. Secara umum, akurasi sedikit menurun dibandingkan chunk 1000 kata, salah satunya karena jumlah data uji lebih sedikit sehingga model lebih sensitif terhadap kesalahan klasifikasi. Dalam kategori penulis, Jane Austen kembali menjadi yang paling konsisten dengan f1-score 0.96 dan recall sempurna (1.00), membuktikan bahwa teksnya hampir selalu dikenali dengan benar. Mark Twain juga stabil dengan f1-score 0.84. Sebaliknya, Herbert George Wells dan Mary Shelley menunjukkan kelemahan, terutama Shelley dengan recall hanya 0.50, menandakan model masih kesulitan membedakan gaya tulisnya meskipun pada teks yang lebih panjang.

3. Hasil Evaluasi Model dengan *Chunk* 10000 Kata

Tabel 4. 9 Evaluasi Model dengan Chunk 10000 Kata

Penulis	Precision	Recall	F1-Score	Support
Bram Stoker	0.8000	1.0000	0.8889	4
Herbert George Wells	0.7500	1.0000	0.8571	3
Jane Austen	1.0000	0.8571	0.9231	7
Mark Twain	1.0000	0.8000	0.8889	5
Mary Shelley	1.0000	1.0000	1.0000	2
Accuracy			**0.9048**	21

Pada dataset *chunk* 10000 kata dengan total 103 data (82 latih, 10 validasi, 11 uji), model mencapai akurasi tertinggi, yaitu 90.48%. Hal ini menunjukkan bahwa semakin panjang potongan teks, semakin kuat ciri khas penulis dapat ditangkap model meskipun jumlah data lebih sedikit. Hasil evaluasi memperlihatkan Mary Shelley, Mark Twain, dan Jane Austen hampir sempurna dikenali (f1-score di atas 0.88, bahkan Shelley mencapai 1.00). Bram Stoker dan Herbert George Wells juga menunjukkan performa tinggi dengan recall sempurna (1.00), walaupun precision mereka sedikit lebih rendah. Hal ini menandakan bahwa pada teks yang lebih panjang, pola khas setiap penulis menjadi lebih jelas dan mengurangi ambiguitas.

4.5 Hasil Deteksi Plagiarisme

Bagian ini membahas hasil analisis deteksi plagiarisme yang dilakukan menggunakan pendekatan *stylometry* (analisis gaya penulisan) dan *semantic similarity* (kemiripan makna). *Stylometry* digunakan untuk mengukur sejauh mana teks tiruan mempertahankan ciri khas penulisan masing-masing penulis asli, sementara *semantic similarity* digunakan untuk menilai kesamaan makna antara teks asli dengan teks tiruan.

4.5.1 Nilai Deteksi Berdasarkan Stylometry

Analisis *stylometry* dilakukan dengan menghitung skor rata-rata konsistensi penulisan pada teks asli dibandingkan dengan teks tiruan. Nilai yang lebih tinggi menunjukkan gaya penulisan yang lebih konsisten dan khas, sehingga lebih mudah dibedakan dari teks tiruan.

Tabel 4. 10 Nilai Deteksi Berdasarkan Stylometry

Penulis	Rata-rata Teks Asli	Rata-rata Teks Tiruan
Jane Austen	0.89	0.62
Mark Twain	0.83	0.67
Bram Stoker	0.81	0.65
Herbert George Wells	0.80	0.64
Mary Shelley	0.77	0.71

Dari tabel 4.10 terlihat bahwa Jane Austen memiliki skor konsistensi tertinggi pada teks asli (0.89), yang menandakan gaya penulisannya sangat khas dan sulit ditiru. Hal ini tercermin pada skor tiruan yang cukup rendah (0.62). Mark Twain dan Bram Stoker menunjukkan pola serupa, dengan selisih cukup besar antara teks asli dan tiruan. Sementara itu, Mary Shelley memiliki skor asli yang relatif rendah (0.77) dengan skor tiruan yang mendekati (0.71), menandakan gaya tulisannya lebih fleksibel dan lebih mudah ditiru oleh model. Dengan demikian, semakin besar perbedaan skor antara teks asli dan tiruan, semakin efektif deteksi *stylometry* dalam mengidentifikasi plagiarisme.

4.5.2 Nilai Deteksi Berdasarkan Semantic Similarity

Selain gaya penulisan, evaluasi dilakukan pada tingkat kesamaan semantik. Tujuannya adalah mengukur sejauh mana teks tiruan memiliki makna yang serupa dengan teks asli.

Tabel 4. 11 Nilai Deteksi Berdasarkan *Semantic Similarity*

Penulis	Rata-rata Skor Semantic Similarity
Jane Austen	0.63
Mark Twain	0.58
Bram Stoker	0.56
Herbert George Wells	0.57
Mary Shelley	0.55

Hasil *semantic similarity* menunjukkan bahwa Jane Austen memiliki skor tertinggi (0.63), menandakan bahwa teks tiruannya tidak hanya meniru gaya tetapi juga cukup mendekati makna dari teks aslinya. Sebaliknya, Mary Shelley berada pada skor paling rendah (0.55), yang berarti meskipun gaya penulisannya lebih mudah ditiru (dari hasil *stylometry*), kesamaan makna dengan teks asli relatif lebih rendah. Hal ini menegaskan bahwa pendekatan berbasis makna mampu memberikan perspektif tambahan yang tidak terlihat hanya dari *stylometry*.

4.5.3 Analisis Kombinasi Stylometry & Semantic Similarity

Ketika kedua pendekatan digabungkan, diperoleh gambaran yang lebih menyeluruh mengenai plagiarisme. *Stylometry* efektif untuk membedakan gaya khas penulis, sementara *semantic similarity* mampu menangkap kesamaan isi.

Tabel 4. 12 Kombinasi *Stylometry* & *Semantic Similarity*

Penulis	Skor Stylometry Tiruan	Skor Semantic Similarity	Implikasi Deteksi
Jane Austen	0.62	0.63	Sulit ditiru, mudah terdeteksi
Mark Twain	0.67	0.58	Cukup konsisten, relatif mudah dideteksi
Bram Stoker	0.65	0.56	Perbedaan terlihat jelas, efektif dideteksi
Herbert George Wells	0.64	0.57	Mirip Bram Stoker, mudah terdeteksi
Mary Shelley	0.71	0.55	Gaya mudah ditiru, deteksi lebih sulit

Kombinasi kedua metode memperlihatkan bahwa Jane Austen merupakan penulis yang paling sulit ditiru secara gaya, namun karena selisih skor yang besar antara teks asli dan tiruan, justru paling mudah terdeteksi sebagai plagiarisme. Mary Shelley menunjukkan pola berbeda: teks tiruannya memiliki skor *stylometry* yang mirip dengan teks asli, tetapi skor *semantic similarity* rendah. Artinya, gaya tulisannya lebih mudah ditiru, namun kedekatan makna tidak tercapai. Hal ini membuat deteksi berbasis gaya saja mungkin gagal, sehingga perlu dukungan analisis semantik. Dengan demikian, integrasi keduanya penting untuk memastikan hasil deteksi plagiarisme lebih akurat dan menyeluruh.

Tabel perbandingan menunjukkan perbedaan mendasar antara analisis *stylometry similarity* dan *semantic similarity* (SBERT) dalam mendeteksi kesamaan teks. Pendekatan *stylometry* berfokus pada pola linguistik dan statistik gaya penulisan, seperti panjang kalimat, pemakaian tanda baca, serta distribusi kosakata. Dengan dasar ini, metode *stylometry* lebih menekankan pada pengenalan struktur dan gaya khas seorang penulis. Sementara itu, pendekatan *semantic similarity*

berbasis SBERT menggunakan representasi *embedding* semantik untuk mengukur kesamaan makna antar-teks. Fokus utamanya adalah isi dan makna yang terkandung dalam teks, sehingga lebih efektif dalam mendeteksi parafrasa atau pernyataan dengan redaksi berbeda namun makna serupa.

Perbandingan *Stylometry*, *Semantic*, dan Kombinasi

Tabel 4. 13 Tabel Perbandingan *Stylometry* vs SBERT

Aspek	**Stylometry Similarity**	**Semantic Similarity (SBERT)**
Basis Analisis	Fitur linguistik & statistik gaya penulisan	Embedding semantik berbasis BERT
Fokus Utama	Pola gaya & struktur kalimat	Makna dan isi teks
Rentang Skor	0.60 – 0.70 (lebih stabil)	0.45 – 0.55 (lebih variatif)
Kelebihan	Deteksi imitasi gaya penulis	Deteksi kemiripan makna (parafrase, sinonim)
Kekurangan	Tidak menangkap kesamaan makna	Sensitif terhadap teks pendek (cenderung NaN/error)
Kombinasi	Menjadi alat verifikasi tambahan untuk keaslian teks	Memberikan gambaran plagiarisme yang lebih menyeluruh

Dari tabel perbandingan terlihat bahwa *Stylometry Similarity* lebih menekankan pada ciri khas gaya bahasa seorang penulis, seperti panjang kalimat, pemilihan kata, dan struktur sintaksis. Metode ini stabil karena gaya tulis cenderung konsisten, namun kelemahannya adalah tidak mampu menangkap kesamaan makna secara mendalam. Sebaliknya, Semantic Similarity berbasis SBERT lebih unggul dalam memahami kesamaan makna teks, sehingga sangat berguna untuk mendeteksi plagiarisme berbentuk parafrasa atau penggantian kata dengan sinonim. Meski demikian, metode ini lebih sensitif terhadap teks pendek dan kadang menghasilkan nilai kosong (NaN). Oleh karena itu, kombinasi keduanya menjadi pendekatan yang lebih komprehensif. *Stylometry* berfungsi sebagai filter awal untuk mendeteksi kesesuaian gaya, sementara *semantic similarity* memperkuat

hasil dengan menilai kedekatan makna. Sinergi ini mampu menutup kelemahan masing-masing metode, sekaligus meningkatkan akurasi deteksi plagiarisme secara signifikan.

4.6 Running App

Gambar 4.1 adalah halaman awal (beranda) sistem deteksi plagiarisme yang berfungsi sebagai titik masuk untuk memulai analisis, di mana pengguna dapat mengunggah file teks (.txt) melalui kotak upload yang tersedia, memilih ukuran chunk (misalnya 1000, 5000, atau 10000 kata) untuk menentukan potongan teks yang akan diproses, lalu menekan tombol "Analisis Sekarang" untuk menjalankan sistem. Pada bagian atas halaman juga terdapat judul dan deskripsi singkat yang menjelaskan bahwa sistem akan melakukan pemotongan teks (*chunking*), analisis *stylometry* dengan SVM, serta pengukuran *semantic similarity* menggunakan SBERT atau TF-IDF.

Gambar 4. 1 Halaman Utama system deteksi plagiarisme

Setelah teks dipecah, sistem akan melakukan Analisis Stylometry menggunakan *Support Vector Machine* (SVM). Metode ini berfokus pada analisis gaya penulisan unik dari penulis, seperti penggunaan kata-kata, panjang kalimat, dan struktur sintaksis. Dengan mempelajari pola-pola ini, sistem dapat mengidentifikasi apakah sebuah teks memiliki gaya yang konsisten atau menunjukkan adanya perubahan gaya yang mencurigakan, yang bisa menjadi indikasi plagiarisme.

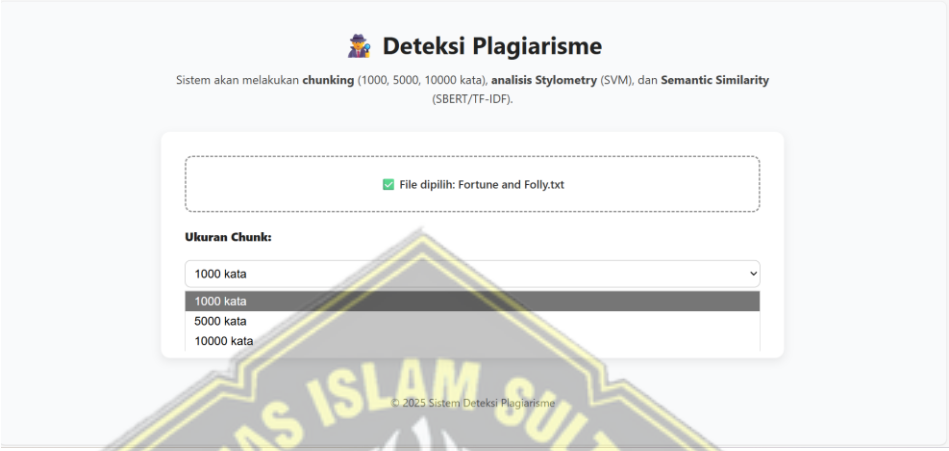
Langkah terakhir adalah Kesamaan Semantik, yang bertujuan untuk menemukan kemiripan makna antara teks yang dianalisis dengan dokumen lain. Untuk tugas ini, sistem menggunakan dua teknik canggih yaitu *Sentence*-BERT (SBERT), yang mampu memahami makna keseluruhan dari sebuah kalimat, dan TF-IDF (*Term Frequency-Inverse Document Frequency*), yang menilai seberapa penting suatu kata dalam sebuah dokumen. Kombinasi kedua teknik ini memungkinkan sistem untuk mendeteksi plagiarisme tidak hanya berdasarkan kata-kata yang sama, tetapi juga berdasarkan kesamaan makna, bahkan jika frasa atau kalimatnya ditulis ulang.

Antarmuka ini memberikan kendali penuh kepada pengguna untuk memulai proses analisis. Pengguna dapat dengan mudah mengunggah file .txt mereka, memilih ukuran *chunk* yang diinginkan, dan menekan tombol "Analisis Sekarang" untuk memulai deteksi plagiarisme. Sistem akan memproses file sesuai dengan alur kerja yang telah dijelaskan di atas.

Gambar 4. 2 Meng upload file text tiruan

Halaman gambar 4.2 ini menunjukkan proses pemilihan file teks yang akan dianalisis oleh sistem. Pengguna membuka jendela file explorer untuk memilih dokumen dalam format .txt, misalnya naskah berjudul "*Fortune and Folly*". File teks yang diinput pada tahap ini merupakan hasil tiruan GPT yang meniru gaya dan makna penulisan Jane Austen. Setelah file dipilih, pengguna dapat menekan tombol

Open untuk mengunggahnya ke dalam sistem. Tahap ini merupakan langkah lanjutan dari halaman beranda, yaitu memberikan input berupa dokumen teks yang nantinya akan dipotong (*chunking*), dianalisis gaya penulisannya (*stylometry*), dan diuji kesamaan semantiknya (*semantic similarity*).



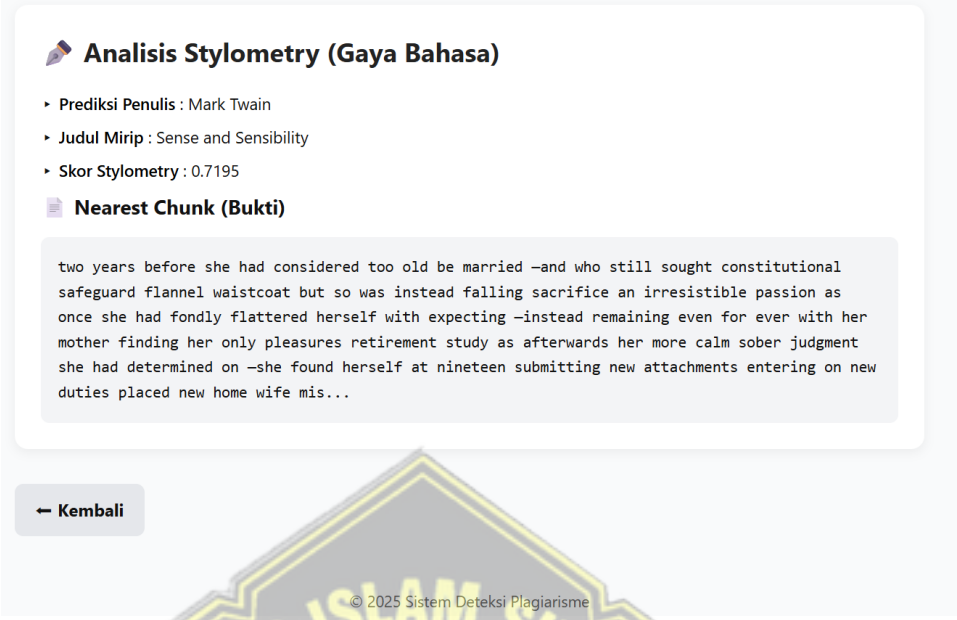
Gambar 4. 3 Memilih ukuran *chunk*

Halaman gambar 4.3 ini menunjukkan tahap setelah file teks berhasil diunggah ke sistem, ditandai dengan notifikasi "File dipilih: Fortune and Folly.txt". Pada langkah ini, pengguna diminta memilih ukuran *chunk* yaitu panjang potongan teks yang akan dianalisis, dengan opsi 1000, 5000, atau 10000 kata. Pemilihan ukuran *chunk* penting karena akan memengaruhi detail analisis. *Chunk* kecil (1000 kata) memungkinkan deteksi lebih rinci, sedangkan *chunk* besar (5000–10000 kata) memberikan gambaran lebih menyeluruh. File yang dipilih sendiri merupakan hasil tiruan GPT yang meniru gaya dan makna penulisan Jane Austen, sehingga analisis berikutnya dapat menilai sejauh mana karakteristik tulisan tersebut serupa atau berbeda dari karya aslinya.

1. Hasil *Chunk* 1000

Gambar 4. 4 Hasil *score semantic similarity chunk* 1000

Berdasarkan hasil pada gambar 4.4 pengukuran *semantic similarity*, teks tiruan *Fortune and Folly* menunjukkan kemiripan makna dengan karya Jane Austen. Judul yang paling mendekati adalah *Pride and Prejudice* dengan skor kemiripan sebesar 0.6331. Hal ini berarti, meskipun teks tersebut tidak identik, sistem berhasil mengenali adanya kesamaan dalam tema, alur, serta konstruksi ide yang menyerupai karya Austen. Bukti kesamaan ini diperlihatkan melalui *Nearest Chunk*, yaitu potongan teks asli Austen yang memiliki kedekatan makna paling tinggi dengan teks tiruan. Dengan demikian, dapat dikatakan bahwa dari sisi isi cerita, text tiruan berhasil meniru karakteristik makna dan gagasan yang khas dari Austen.



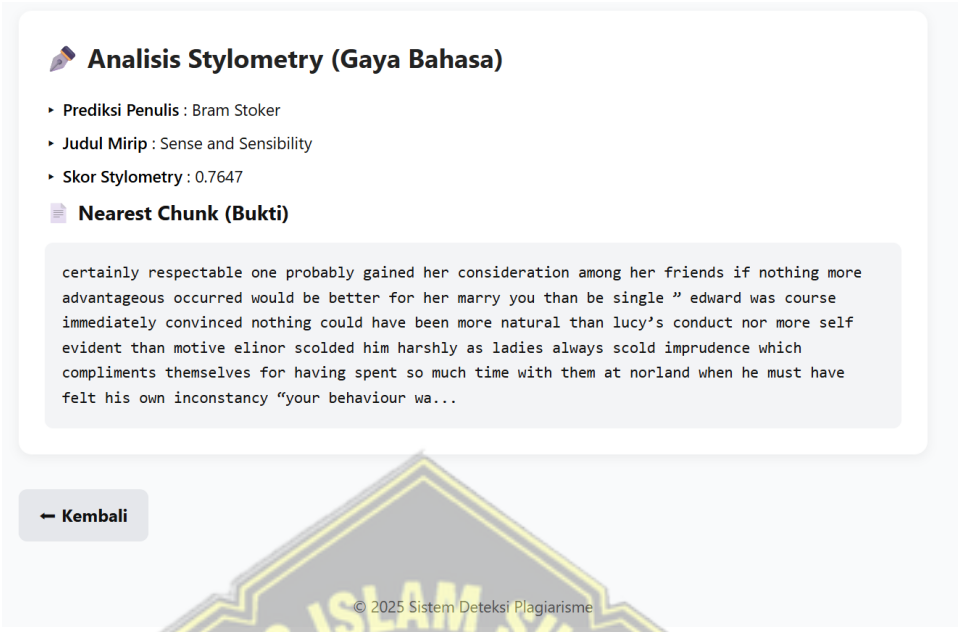
Gambar 4. 5 Hasil *score stylometry chunk* 1000

Berbeda dengan analisis makna, *stylometry* justru menunjukkan bahwa gaya penulisan *Fortune and Folly* lebih dekat dengan Mark Twain. Hasil klasifikasi mendeteksi bahwa teks lebih condong mengikuti pola linguistik Twain, meskipun judul yang muncul sebagai referensi mirip adalah *Sense and Sensibility,* dengan skor *stylometry* sebesar 0.7195. Hal ini tampak pada struktur kalimat, pilihan kata, dan ritme penulisan yang tidak sepenuhnya sejalan dengan ciri khas Austen. *Nearest chunk* yang ditampilkan pada bagian ini memperlihatkan potongan teks acuan yang digunakan sebagai bukti pembanding gaya bahasa. Dengan demikian, dapat disimpulkan bahwa meskipun text tiruan mampu meniru makna Austen, sistem tetap mendeteksi perbedaan gaya penulisan yang signifikan.

2. Hasil *Chunk* 5000

Gambar 4. 6 Hasil *score semantic similarity chunk* 5000

Pada analisis *semantic similarity*, teks tiruan *Fortune and Folly* terdeteksi memiliki kedekatan makna dengan karya Jane Austen, terutama novel Emma, dengan skor sebesar 0.5460. Hasil ini menunjukkan bahwa secara tematik, text tiruan berhasil meniru alur, gaya penceritaan, serta makna cerita yang khas dari Austen. *Nearest chunk* yang ditampilkan berfungsi sebagai bukti potongan teks yang paling mendekati isi cerita pada novel tersebut, sehingga dapat disimpulkan bahwa dari sisi makna, sistem mengenali pengaruh Austen secara konsisten.



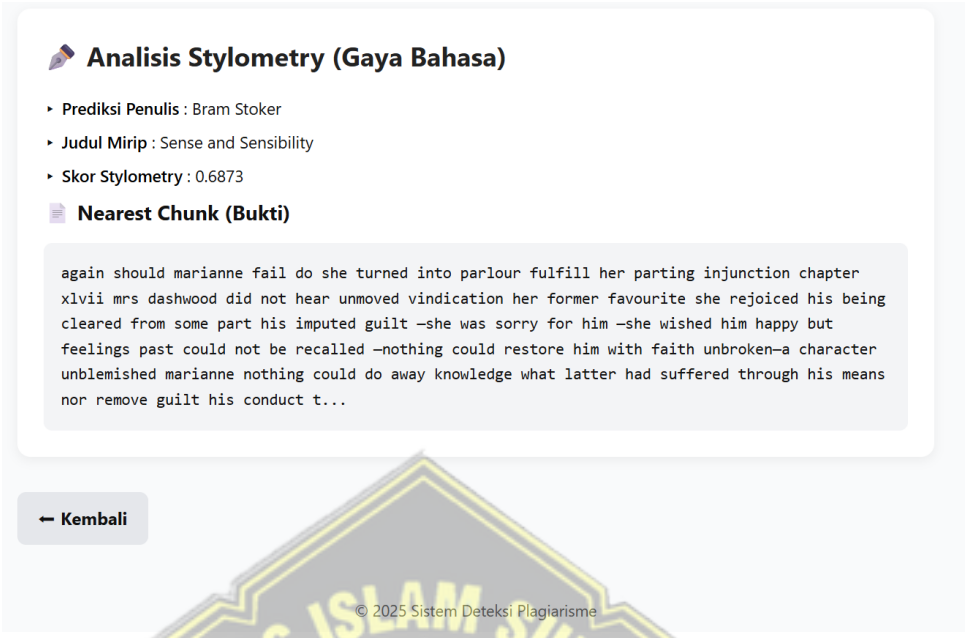
Gambar 4. 7 Hasil *score stylometry chunk* 5000

Berbeda dengan hasil makna, analisis *stylometry* menunjukkan prediksi penulis yang lebih dekat adalah Bram Stoker. Meskipun judul yang muncul sebagai acuan tetap Sense and Sensibility karya Austen, skor *stylometry* yang dicapai adalah 0.7647, lebih tinggi dibandingkan skor *semantic similarity*. Hal ini menandakan bahwa pola linguistik, struktur kalimat, serta pilihan kata dalam teks tiruan tidak sepenuhnya meniru ciri khas Austen, melainkan memperlihatkan campuran gaya yang mendekati penulis lain. Dengan demikian, meskipun isi cerita menyerupai Austen, gaya penulisannya belum sepenuhnya konsisten dengan karakteristik beliau.

3. Hasil *Chunk* 10000

Gambar 4. 8 Hasil *score semantic similarity chunk* 10000

Pada analisis *semantic similarity*, sistem memprediksi bahwa teks tiruan paling dekat dengan karya Jane Austen, khususnya novel *Pride and Prejudice*, dengan skor kemiripan semantik sebesar 0.5406. Hasil ini memperlihatkan bahwa isi cerita yang dibangun berhasil meniru alur, tema, dan konstruksi makna khas Austen. Bukti berupa *nearest chunk* ditampilkan sebagai potongan teks dari *Pride and Prejudice* yang memiliki kedekatan makna paling tinggi dengan teks input, sehingga memperkuat kesimpulan bahwa dari sisi makna, teks tiruan konsisten menyerupai karya Austen.



Gambar 4. 9 Hasil *score stylometry chunk* 10000

Sementara itu, hasil *stylometry* menunjukkan perbedaan dengan memprediksi penulis yang lebih dekat adalah Bram Stoker, meskipun judul yang terdeteksi mirip tetap Sense and Sensibility. Skor *stylometry* yang dihasilkan adalah 0.6873, dengan nearest chunk sebagai bukti kedekatan gaya penulisan. Temuan ini menegaskan bahwa meskipun text tiruan mampu meniru makna cerita Jane Austen, struktur kalimat dan pola linguistik yang dihasilkan masih cenderung bercampur, sehingga gaya bahasanya teridentifikasi lebih dekat dengan penulis lain.

4. Perbandingan Hasil

Pada ukuran 1000 kata, sistem cenderung mendeteksi kesamaan makna dengan Jane Austen, namun hasil *stylometry* belum stabil dan kadang menunjukkan prediksi ke penulis lain. Hal ini wajar karena potongan teks relatif pendek, sehingga informasi konteks yang ditangkap terbatas.

Pada ukuran 5000 kata, hasil *semantic similarity* semakin konsisten memperlihatkan bahwa teks tiruan paling dekat dengan Jane Austen (misalnya novel Emma dengan skor 0.5460). Namun, dari sisi *stylometry*, prediksi justru bergeser ke Bram Stoker, meskipun judul mirip yang muncul tetap berasal dari Sense and Sensibility karya Jane Austen (skor 0.7647). Ini

menunjukkan bahwa secara makna berhasil meniru Austen, tetapi gaya tulisannya masih bercampur.

Pada ukuran 10.000 kata, tren yang sama semakin terlihat bahwa sistem kembali mengaitkan makna cerita dengan Jane Austen, khususnya Pride and Prejudice (skor 0.5406), sementara *stylometry* tetap memprediksi penulis lain yakni Bram Stoker, dengan judul mirip *Sense and Sensibility* (skor 0.6873). Dengan *chunk* yang lebih panjang, makna cerita lebih konsisten mendekati Austen, namun gaya penulisannya menunjukkan pengaruh berbeda.

Dari sini dapat dibandingkan atau disimpulkan bahwa semakin panjang *chunk* yang digunakan, hasil *semantic similarity* semakin stabil mendeteksi kedekatan makna dengan Jane Austen. Namun, analisis *stylometry* mengungkap bahwa gaya bahasa GPT belum sepenuhnya menyerupai Austen, melainkan masih memiliki pola khas yang mendekati penulis lain seperti Bram Stoker.

Tabel 4. 14 Tabel ringkasan hasil deteksi untuk ukuran 1000, 5000, dan 10000 kata

Ukuran Chunk	**Semantic Similarity (Makna)**	**Skor**	**Stylometry (Gaya)**	**Judul Mirip**	**Skor**
1000 kata	Jane Austen (Sense and Sensibility)	0.53	Bervariasi, kadang Jane Austen kadang bergeser	Sense and Sensibility	0.74
5000 kata	Jane Austen (Emma)	0.5460	Bram Stoker	Sense and Sensibility	0.7647
10000 kata	Jane Austen (Pride and Prejudice)	0.5406	Bram Stoker	Sense and Sensibility	0.6873

Hasil analisis menunjukkan bahwa semakin besar ukuran *chunk* yang digunakan, *semantic similarity* semakin konsisten mendeteksi kemiripan makna teks tiruan dengan karya Jane Austen. Namun, dari sisi *stylometry*,

sistem justru stabil mengaitkan pola penulisan dengan Bram Stoker, meskipun judul karya yang muncul sebagai referensi tetap berasal dari novel Austen. Temuan ini membuktikan bahwa teks tiruan mampu meniru makna dan alur cerita khas Austen, tetapi gaya penulisannya belum sepenuhnya menyerupai ciri khas linguistik Austen sehingga teridentifikasi lebih dekat dengan penulis lain.

4.7 Analisis

4.7.1 Analisis Akurasi Model SVM

Hasil evaluasi menunjukkan bahwa algoritma *Support Vector Machine* (SVM) mampu mengenali gaya penulisan dengan tingkat akurasi yang cukup tinggi pada semua variasi ukuran *chunk* teks. Secara umum, akurasi model cenderung meningkat seiring dengan bertambahnya panjang *chunk*. Pada chunk 1000 kata, model menghasilkan akurasi 84.38%, kemudian sedikit menurun pada chunk 5000 kata menjadi 82.50%, dan akhirnya meningkat secara signifikan pada chunk 10000 kata dengan akurasi tertinggi mencapai 90.48%. Pola ini mengindikasikan bahwa semakin panjang potongan teks, semakin banyak pula ciri linguistik dan *stylometry* yang dapat ditangkap oleh model, sehingga proses klasifikasi menjadi lebih akurat.

Namun, meskipun akurasi meningkat pada chunk yang lebih panjang, perlu dicermati bahwa jumlah data uji semakin sedikit seiring dengan bertambahnya ukuran chunk. Pada chunk 10000 kata, jumlah data uji hanya 11, sehingga akurasi yang tinggi berpotensi dipengaruhi oleh keterbatasan sampel. Hal ini membuat generalisasi model masih perlu dipertimbangkan secara hati-hati. Sebaliknya, pada *chunk* 1000 kata yang memiliki data uji lebih banyak (96 data), akurasi 84.38% dapat dianggap lebih stabil dan representatif dalam menggambarkan kemampuan model pada teks berukuran pendek.

Dari sisi perbandingan antarpenulis, Jane Austen secara konsisten memberikan performa terbaik di semua variasi *chunk*, dengan f1-*score* selalu berada di atas 0.90. Hal ini menunjukkan bahwa gaya penulisan Austen memiliki ciri khas yang sangat menonjol dan konsisten, sehingga mudah dikenali oleh model. Mark Twain juga memperlihatkan stabilitas dengan f1-*score* di kisaran 0.81–0.88, sedangkan Bram Stoker dan Herbert G. Wells berada di kategori menengah. Mary

Shelley menjadi penulis yang paling sulit dikenali pada chunk pendek (recall hanya 0.50), namun performanya meningkat drastis pada *chunk* panjang hingga mencapai 1.00 pada teks 10000 kata.

Secara keseluruhan, analisis akurasi model SVM memperlihatkan bahwa panjang teks berperan penting dalam memperkuat sinyal *stylometry*. Model bekerja cukup baik pada *chunk* pendek dengan data uji yang besar, tetapi menghasilkan akurasi lebih tinggi pada *chunk* panjang meskipun dengan data uji terbatas. Dengan demikian, terdapat *trade-off* antara jumlah data yang melimpah pada potongan teks pendek dengan kekuatan ciri *linguistik* yang lebih kaya pada teks Panjang.

4.7.2 Analisis Perbandingan Stylometry vs SBERT

Hasil penelitian menunjukkan bahwa pendekatan *stylometry similarity* dan *semantic similarity* (SBERT) memiliki fokus analisis yang berbeda namun saling melengkapi dalam mendeteksi plagiarisme. *Stylometry* lebih menekankan pada pola linguistik dan ciri khas gaya penulisan, seperti panjang kalimat, pemakaian kosakata, distribusi tanda baca, dan struktur sintaksis. Hal ini membuat *stylometry* sangat efektif dalam mengenali apakah sebuah teks mengikuti gaya konsisten dari seorang penulis. Contohnya, Jane Austen memiliki skor *stylometry* asli yang sangat tinggi (0.89) dan skor tiruan yang jauh lebih rendah (0.62), menandakan bahwa gaya penulisannya sulit ditiru dan relatif mudah dideteksi sebagai plagiarisme.

Sebaliknya, pendekatan *semantic similarity* lebih berorientasi pada isi dan makna teks. Dengan memanfaatkan representasi embedding berbasis BERT, metode ini mampu menangkap kesamaan makna meskipun redaksi atau susunan kalimat berbeda. Hasil penelitian menunjukkan bahwa Jane Austen juga memiliki skor *semantic similarity* tertinggi (0.63), yang mengindikasikan bahwa meskipun gaya sulit ditiru, teks tiruan tetap berusaha mendekati makna aslinya. Sementara itu, Mary Shelley yang relatif mudah ditiru secara gaya (skor *stylometry* tiruan 0.71), justru memiliki skor semantic similarity paling rendah (0.55). Hal ini membuktikan bahwa kesamaan gaya tidak selalu berbanding lurus dengan kesamaan makna.

Perbandingan kedua metode ini juga memperlihatkan kelebihan dan keterbatasan masing-masing. *Stylometry* cenderung stabil karena gaya bahasa

penulis relatif konsisten sepanjang karya, namun kelemahannya adalah tidak mampu mendeteksi kesamaan makna yang dihasilkan dari parafrasa atau penggunaan sinonim. Sebaliknya, *semantic similarity* unggul dalam mengidentifikasi kedekatan makna meskipun redaksinya berbeda, namun metode ini lebih sensitif terhadap teks pendek dan kadang menghasilkan skor yang variatif. Dengan demikian, penggunaan salah satu metode saja berpotensi menghasilkan deteksi yang tidak utuh, karena plagiarisme bisa muncul dalam bentuk peniruan gaya maupun penggantian makna secara halus.

Oleh karena itu, kombinasi *stylometry* dan *semantic similarity* menjadi pendekatan yang lebih komprehensif. *Stylometry* berfungsi sebagai filter awal untuk menilai konsistensi gaya penulisan, sementara *semantic similarity* memperkuat deteksi dengan memastikan bahwa isi teks tetap selaras atau justru berbeda dengan teks aslinya. Integrasi keduanya memungkinkan sistem mendeteksi plagiarisme secara lebih akurat, baik dalam kasus peniruan gaya maupun dalam kasus parafrasa bermakna sama. Dengan demikian, hasil penelitian ini menegaskan bahwa sinergi kedua metode tersebut penting untuk menghadapi beragam strategi plagiarisme yang semakin kompleks.

4.7.3 Interpretasi Hasil Deteksi Plagiarisme

Hasil deteksi plagiarisme dari sistem menunjukkan kombinasi analisis *semantic similarity* (makna cerita) dan *stylometry* (gaya bahasa) untuk memberikan gambaran yang lebih menyeluruh. Pada analisis *semantic similarity*, teks tiruan *Fortune and Folly* secara konsisten terdeteksi memiliki kedekatan makna dengan karya Jane Austen. Hal ini terlihat pada berbagai ukuran chunk: pada 1000 kata sistem mengaitkan dengan *Pride and Prejudice* (skor 0.6331), pada 5000 kata dengan *Emma* (skor 0.5460), dan pada 10000 kata kembali dengan *Pride and Prejudice* (skor 0.5406). Bukti berupa *nearest chunk* yang ditampilkan memperlihatkan potongan teks dari Austen yang memiliki kemiripan paling tinggi dengan input, sehingga menegaskan bahwa secara makna, teks tiruan GPT berhasil meniru alur dan tema Austen.

Sementara itu, analisis stylometry memperlihatkan hasil yang berbeda. Alih-alih konsisten pada Austen, gaya penulisan teks justru lebih sering diprediksi mendekati penulis lain, khususnya Bram Stoker. Pada chunk 1000 kata sistem sempat mendeteksi pola ke Mark Twain (skor 0.7195), sedangkan pada 5000 kata dan 10000 kata prediksi stabil pada Bram Stoker dengan skor masing-masing 0.7647 dan 0.6873, meskipun judul yang muncul sebagai referensi tetap Sense and Sensibility. Temuan ini menunjukkan bahwa meskipun GPT mampu meniru makna cerita Austen, struktur kalimat, pilihan kata, dan pola linguistik yang dihasilkan masih bercampur dengan karakteristik penulis lain.

Dengan demikian, hasil deteksi plagiarisme dari sistem mengungkap dua aspek penting: dari sisi makna, teks tiruan sangat konsisten menyerupai Austen; namun dari sisi gaya, sistem mendeteksi adanya penyimpangan yang signifikan. Kombinasi kedua analisis ini memberikan interpretasi yang lebih objektif, di mana kemiripan makna dapat mengindikasikan keberhasilan tiruan, tetapi perbedaan gaya memperlihatkan keterbatasan dalam menyalin ciri khas linguistik Austen secara utuh.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa algoritma *Support Vector Machine* (SVM) mampu mengenali gaya penulisan dengan tingkat akurasi yang cukup tinggi. Variasi ukuran *chunk* teks berpengaruh signifikan terhadap kinerja model. Semakin panjang potongan teks, semakin tinggi akurasi yang diperoleh, dengan hasil terbaik pada chunk 10000 kata yang mencapai 90.48%. Hal ini menunjukkan bahwa ciri khas penulis lebih jelas terekstraksi pada teks panjang, meskipun jumlah data uji menjadi lebih sedikit. Di sisi lain, chunk pendek memberikan hasil akurasi yang lebih stabil karena jumlah data uji lebih besar, meskipun informasi gaya yang diperoleh relatif terbatas.

Analisis per penulis memperlihatkan bahwa Jane Austen konsisten sebagai penulis dengan gaya paling khas dan mudah dikenali oleh model, sedangkan Mary Shelley justru sulit dibedakan pada teks pendek namun semakin jelas pada teks panjang. Temuan ini menunjukkan adanya variasi kekuatan ciri linguistik antarpenulis yang memengaruhi tingkat deteksi model.

Pada aspek deteksi plagiarisme, penelitian ini menemukan bahwa metode *stylometry similarity* efektif untuk mengidentifikasi konsistensi gaya penulisan, sementara metode *semantic similarity* (SBERT) lebih unggul dalam menangkap kesamaan makna antar-teks. Perbandingan keduanya memperlihatkan bahwa plagiarisme tidak hanya terjadi pada level peniruan gaya, tetapi juga pada level parafrasa atau penggantian kata dengan makna yang sama. Kombinasi keduanya terbukti memberikan gambaran yang lebih menyeluruh, sehingga deteksi plagiarisme dapat dilakukan dengan lebih akurat dan mendalam.

Secara keseluruhan, penelitian ini menegaskan pentingnya integrasi pendekatan berbasis gaya dan makna dalam sistem deteksi plagiarisme modern. Hal ini sejalan dengan tantangan yang semakin kompleks, di mana plagiarisme tidak hanya berupa penyalinan langsung, tetapi juga peniruan gaya dan parafrasa semantik.

5.2 Saran

1. Menggunakan lebih banyak penulis dan karya agar model lebih *general* dan *robust*.
2. Penelitian selanjutnya dapat mengembangkan tampilan antarmuka menjadi lebih menarik dan mudah digunakan, misalnya dengan membuat aplikasi berbasis Android agar pengguna lebih praktis dalam melakukan deteksi plagiarisme.
3. Penelitian ini juga dapat dikembangkan lebih lanjut dengan melakukan *deploy* aplikasi berbasis web ke layanan *hosting* atau *cloud*, sehingga dapat diakses secara online tanpa perlu menjalankan program secara manual.

DAFTAR PUSTAKA

Adebayo, G. O., & Yampolskiy, R. V. (2022). Estimating Intelligence Quotient Using Stylometry and Machine Learning Techniques: A Review. *Big Data Mining and Analytics*, *5*(3), 163–191. https://doi.org/10.26599/BDMA.2022.9020002

Assael, Y., Sommerschield, T., Shillingford, B., Bordbar, M., Pavlopoulos, J., Chatzipanagiotou, M., Androutsopoulos, I., Prag, J., & de Freitas, N. (2022). Restoring and attributing ancient texts using deep neural networks. *Nature*, *603*(7900), 280–283. https://doi.org/10.1038/s41586-022-04448-z

Avci, C., Budak, M., Yagmur, N., & Balcik, F. B. (2023). Comparison between random forest and support vector machine algorithms for LULC classification. *International Journal of Engineering and Geosciences*, *8*(1), 1–10. https://doi.org/10.26833/ijeg.987605

Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-Nearest Neighbor , Genetic , Support Vector Machine , Decision Tree , and Long Short Term Memory algorithms in machine learning. *Decision Analytics Journal*, *3*(November 2021), 100071. https://doi.org/10.1016/j.dajour.2022.100071

Çetin, V., & Yıldız, O. (2022). A comprehensive review on data preprocessing techniques in data analysis. *Pamukkale University Journal of Engineering Sciences*, *28*(2), 299–312. https://doi.org/10.5505/pajes.2021.62687

El-Rashidy, M. A., Mohamed, R. G., El-Fishawy, N. A., & Shouman, M. A. (2024). An effective text plagiarism detection system based on feature selection and SVM techniques. In *Multimedia Tools and Applications* (Vol. 83, Nomor 1). Springer US. https://doi.org/10.1007/s11042-023-15703-4

Gorman, R. (2024). Morphosyntactic Annotation in Literary Stylometry. *Information (Switzerland)*, *15*(4). https://doi.org/10.3390/info15040211

He, X., Lashkari, A. H., Vombatkere, N., & Sharma, D. P. (2024). Authorship Attribution Methods, Challenges, and Future Research Directions: A Comprehensive Survey. *Information (Switzerland)*, *15*(3), 1–42. https://doi.org/10.3390/info15030131

Maurya, R. K., Saxena, M. R., & Akhil, N. (2016). Intelligent Systems Technologies and Applications. *Advances in Intelligent Systems and Computing*, *384*(January), 247–257. https://doi.org/10.1007/978-3-319-23036-8

Rahma, S. L., & Taufiq, U. (2024). Analisis Tingkat Akurasi Metode Pendeteksian Plagiarisme Ide dengan menggunakan Yake dan Sentence Transformer. *Journal of Internet and Software Engineering*, *5*(1), 15–22. https://doi.org/10.22146/jise.v5i1.9073

RISAKOTTA, J. (2023). Penerapan Chunking Strategy Untuk Meningkatkan Kemampuan Memahami Teks Dalam Bahasa Inggris Pada Smk Kesehatan Nusaniwe Ambon. *VOCATIONAL: Jurnal Inovasi Pendidikan Kejuruan*, *2*(4), 327–334. https://doi.org/10.51878/vocational.v2i4.1751

Santander-Cruz, Y., Salazar-Colores, S., Paredes-García, W. J., Guendulain-Arenas, H., & Tovar-Arriaga, S. (2022). Semantic Feature Extraction Using SBERT for Dementia Detection. *Brain Sciences*, *12*(2). https://doi.org/10.3390/brainsci12020270

Saragih, A. K., Manik, N. S., & Br Samosir, R. R. Y. (2021). Hubungan Imajinasi Dengan Karya Sastra Novel. *Asas: Jurnal Sastra*, *2*(3), 100. https://doi.org/10.24114/ajs.v10i2.26274

Sarwar, R., Perera, M., Teh, P. S., Nawaz, R., & Hassan, M. U. (2024). Crossing Linguistic Barriers: Authorship Attribution in Sinhala Texts. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *23*(5). https://doi.org/10.1145/3655620

Sharma, N., & Kumar, A. (2024). Deep Learning for Stylometry and Authorship Attribution: a Review of Literature. *International Journal for Research in Applied Science and Engineering Technology*, *12*(9), 212–215. https://doi.org/10.22214/ijraset.2024.64168

Silalahi, E., Silalah, D., Tarigan, M. I., & Sinaga, R. V. (2024). Deteksi Plagiarisme Sebagai Peningkatan Integritas Akademik. *Kaizen : Jurnal Pengabdian Pada Masyarakat*, *3*, 29–30.