JURNAL ILMIAH SULTAN AGUNG
Universitas Islam Sultan Agung
Semarang, 13 Desember 2025
ISSN: 2963-2730

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

[1]Mey Rini Rz*, [2]Badieah

[1,2]Teknik Informatika Fakultas Teknologi Industri, Universitas Islam Sultan Agung

*Corresponding Author:
meyrini202@gmail.com

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*

1. PENDAHULUAN

Plagiarisme merupakan tindakan tidak etis yang melibatkan pengambilan karya orang lain tanpa memberikan atribusi yang semestinya(Adebayo & Yampolskiy, 2022). 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(Assael et al., 2022). 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 et al., 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 et al., 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.

Penelitian ini bertujuan untuk mengidentifikasi penulis (author) dari potongan teks novel berbahasa Inggris menggunakan analisis gaya penulisan (stylometry) dengan algoritma *Support Vector Machine* (SVM) serta mengembangkan sistem deteksi plagiarisme berdasarkan tingkat kemiripan gaya penulisan dan makna menggunakan model *Sentence-BERT* (SBERT). Melalui pendekatan ini, sistem yang dikembangkan diharapkan mampu mendeteksi indikasi plagiarisme secara otomatis pada teks panjang dengan mengukur kesesuaian antara teks analisis dan karakteristik penulis aslinya. Penerapan metode stylometry dan SVM dalam konteks deteksi plagiarisme karya sastra fiksi membuka peluang baru bagi pengawasan keaslian tulisan, sekaligus memberikan kontribusi penting bagi akademisi, penerbit, dan penulis dalam menjaga orisinalitas serta integritas karya sastra di era digital yang semakin maju dan kompleks.

2. METODE

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(Avci et al., 2023). 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:

1) Studi Literatur

Penelitian ini diawali dengan studi literatur untuk memperdalam pemahaman mengenai konsep *stylometry*, teknik pengolahan teks, serta penerapan algoritma *Support Vector Machine (SVM)* dan model *Sentence-BERT (SBERT)*. Sumber literatur yang digunakan meliputi jurnal ilmiah, artikel, makalah konferensi, dan repositori digital resmi. Tahap ini bertujuan membangun dasar teoritis yang kuat dalam merancang sistem atribusi penulis dan deteksi plagiarisme berbasis pembelajaran mesin. Selain itu, tinjauan dilakukan untuk mengidentifikasi metode ekstraksi fitur linguistik dan pengukuran kesamaan semantik yang relevan. Hasil studi literatur menjadi pedoman utama dalam penyusunan metodologi dan perancangan sistem secara keseluruhan.

2) Pengumpulan Data

Tahap selanjutnya adalah pengumpulan dan persiapan data. Data utama berupa teks novel asli diperoleh dari situs *Project Gutenberg*, sedangkan data tiruan dihasilkan oleh model bahasa seperti GPT untuk meniru gaya dan makna penulis asli. Kedua jenis data ini digunakan untuk proses *authorship attribution* dan deteksi plagiarisme. Teks disimpan dalam format *.txt* agar mudah diproses pada tahap *preprocessing* dan *chunking*. Setiap novel dibagi menjadi potongan teks berukuran 1000, 5000, dan 10000 kata untuk mempermudah analisis fitur serta evaluasi kinerja model (He et al., 2024). Proses ini memastikan bahwa dataset yang digunakan representatif dan terdistribusi secara seimbang antar penulis.

3) Perancangan Sistem

Tahap pemrosesan teks mencakup pembersihan data melalui *preprocessing* seperti penghapusan tanda baca, angka, dan karakter khusus, serta normalisasi huruf dan tokenisasi. Setelah itu dilakukan ekstraksi fitur *stylometry* berupa panjang kalimat, panjang kata, *type-token ratio*, dan distribusi *part-of-speech tag*. Fitur-fitur ini digunakan dalam model *SVM* untuk memprediksi penulis teks berdasarkan pola linguistik. Analisis *semantic similarity* dilakukan menggunakan dua pendekatan, yaitu *TF-IDF* sebagai metode berbasis kata dan *SBERT* sebagai metode berbasis makna(Sharma & Kumar, 2024). Pendekatan ganda ini memungkinkan sistem memahami baik gaya penulisan maupun kedekatan semantik antar teks.

4) Integrasi dan Evaluasi Sistem

Tahap terakhir adalah integrasi hasil analisis *stylometry* dan *semantic similarity* untuk mendeteksi plagiarisme secara komprehensif. Model *SVM* menghasilkan prediksi penulis, sedangkan skor kesamaan semantik menunjukkan tingkat kemiripan makna

antar teks. Evaluasi dilakukan menggunakan metrik akurasi, presisi, *recall*, dan *F1-score* untuk menilai performa model klasifikasi serta efektivitas deteksi plagiarisme(Rahma & Taufiq, 2024). Sistem antarmuka dirancang agar pengguna dapat mengunggah file teks, memilih ukuran *chunk*, dan memperoleh hasil analisis berupa prediksi penulis serta skor kesamaan teks. Dengan pendekatan ini, sistem mampu mendeteksi plagiarisme berbasis gaya dan makna secara akurat serta mendukung upaya menjaga keaslian karya sastra digital.

3. HASIL DAN PEMBAHASAN

3.1 Deskripsi Data Penelitian

Berdasarkan hasil proses *chunking* terhadap lima penulis, diperoleh total 959 potongan teks untuk ukuran 1000 kata, 199 potongan untuk 5000 kata, dan 103 potongan untuk 10000 kata. Penulis dengan jumlah *chunk* terbanyak adalah Jane Austen, diikuti oleh Mark Twain, Bram Stoker, Herbert George Wells, dan Mary Shelley. Variasi jumlah *chunk* ini menunjukkan bahwa panjang novel asli memengaruhi banyaknya data yang dihasilkan, di mana karya yang lebih panjang menghasilkan lebih banyak potongan teks. Setelah seluruh *chunk* dikumpulkan, data kemudian dibagi menjadi tiga bagian, yaitu 80% untuk pelatihan, 10% untuk validasi, dan 10% untuk pengujian, dengan tetap menjaga keseimbangan distribusi antar penulis. Pada dataset 1000 kata, pembagian menghasilkan 767 data latih, 96 data validasi, dan 96 data uji, sedangkan pada dataset 5000 dan 10000 kata masing-masing menghasilkan 159–20–20 dan 82–10–11 data. Pembagian ini memastikan model *machine learning* memiliki data yang cukup untuk belajar, menyesuaikan parameter, dan dievaluasi secara objektif tanpa bias, sehingga hasil klasifikasi penulis dan deteksi plagiarisme dapat dilakukan secara akurat dan seimbang. 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.

3.2 Hasil Implementasi Sistem

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.

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.

3) Hasil Ekstraksi Fitur Stylometry

Proses ekstraksi fitur *stylometry* dilakukan untuk mengidentifikasi ciri khas gaya penulisan tiap penulis melalui analisis potongan teks hasil *chunking* berukuran 1000, 5000, dan 10000 kata. Fitur yang diekstraksi mencakup panjang rata-rata kalimat dan kata, rasio kosakata unik (*type token ratio* dan *hapax legomena*), proporsi huruf vokal, tanda baca, kata umum, kata fungsi, serta kompleksitas karakter (*char n-gram entropy*). Hasil menunjukkan bahwa variasi kosakata lebih tinggi pada potongan pendek dengan nilai *TTR* dan *hapax* besar, sedangkan pada potongan panjang nilainya menurun akibat pengulangan kata yang lebih sering. Sebaliknya, fitur struktural seperti penggunaan *stopword*, *function word*, dan panjang kata relatif stabil di semua ukuran *chunk*. Pola ini menegaskan bahwa gaya penulisan penulis lebih konsisten tercermin melalui struktur dan penggunaan kata fungsi daripada variasi kosakata semata.

Tabel 1. 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 2. 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

Tabel 1 menampilkan tiga dataset hasil ekstraksi fitur *stylometry* berdasarkan ukuran potongan teks (1000, 5000, dan 10000 kata) yang masing-masing mencakup 14 fitur linguistik seperti panjang kalimat, panjang kata, variasi kosakata, penggunaan *stopword* dan *function word*, serta kompleksitas karakter (*char n-gram entropy*). Dataset 1000 kata memberikan detail paling tinggi, sedangkan 5000 dan 10000 kata menampilkan gambaran yang lebih umum terhadap gaya penulis. Berdasarkan Tabel 2, nilai rata-rata *stylometry* berbeda pada tiap penulis tiruan, dengan Herbert George Wells menunjukkan konsistensi tertinggi (0.918 pada chunk 1000 kata), sementara Mary Shelley memiliki stabilitas terendah (0.679 pada chunk 10000 kata). Tren umum menunjukkan bahwa semakin besar ukuran chunk, skor *stylometry* cenderung menurun, yang menandakan meningkatnya variasi gaya dalam potongan teks yang lebih panjang.

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 3. 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	0.469

Tabel 3 menunjukkan bahwa skor *semantic similarity* rata-rata teks tiruan berkisar antara 0.49 hingga 0.63, dengan kecenderungan menurun seiring bertambahnya ukuran *chunk*. Jane Austen (tiruan) memiliki skor tertinggi sebesar 0.639 pada *chunk* 1000 kata, menunjukkan konsistensi makna yang baik, sedangkan Mary Shelley (tiruan) memiliki skor terendah 0.496 pada *chunk* 10000 kata, menandakan kesulitan dalam mempertahankan kesamaan semantik. Pada *chunk* kecil (1000 kata), model GPT mampu menghasilkan teks dengan makna yang cukup dekat dengan karya asli, namun pada *chunk* yang lebih panjang (5000–10000 kata), kesamaan makna menurun akibat meningkatnya variasi naratif dan konteks. Secara umum, Jane Austen dan Herbert G. Wells menunjukkan kestabilan semantik lebih tinggi, sementara Mark Twain dan Mary Shelley cenderung lebih fluktuatif, menggambarkan perbedaan kompleksitas struktur bahasa dan gaya penulisan antarpenulis.

5) Integrasi Stylometry & Semantic Similarity

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

Tabel 4. 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
HerbertGeorge 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, 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.

3.3 Hasil Evaluasi Model Klasifikasi

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*.

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.

Tabel 5. 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.

Tabel 6. 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), model mencapai akurasi 82.5%, sedikit menurun dibanding *chunk* 1000 kata karena jumlah data uji yang lebih kecil. Jane Austen tetap paling konsisten dengan *f1-score* 0.96 dan *recall* 1.00, menunjukkan teksnya hampir selalu teridentifikasi dengan benar, sementara Mark Twain juga stabil dengan *f1-score* 0.84. Sebaliknya, Herbert George Wells dan terutama Mary Shelley menunjukkan performa lebih rendah, dengan *recall* Shelley hanya 0.50, menandakan model masih kesulitan mengenali ciri khas gaya tulisnya pada teks yang lebih panjang.

Tabel 7. 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.

3.4 Hasil Deteksi Plagiarisme

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 8. 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 8 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.

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 9. 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*.

3) Analisis Kombinasi Stylometry & Semantik 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 10. 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 metode *stylometry* dan *semantic similarity* menunjukkan bahwa Jane Austen paling sulit ditiru dari segi gaya, namun selisih skor besar antara teks asli dan tiruan membuatnya mudah terdeteksi sebagai plagiarisme. Sebaliknya, Mary Shelley memiliki gaya yang mudah ditiru, tetapi skor *semantic similarity*-nya rendah, menandakan kemiripan makna tidak tercapai. Hal ini menunjukkan bahwa deteksi berbasis gaya saja tidak cukup dan perlu dilengkapi analisis semantik. Pendekatan

stylometry berfokus pada pola linguistik seperti panjang kalimat, tanda baca, dan distribusi kosakata, sedangkan *semantic similarity* berbasis *SBERT* menilai kesamaan makna melalui representasi semantik. Integrasi keduanya penting untuk menghasilkan deteksi plagiarisme yang lebih akurat dan komprehensif.

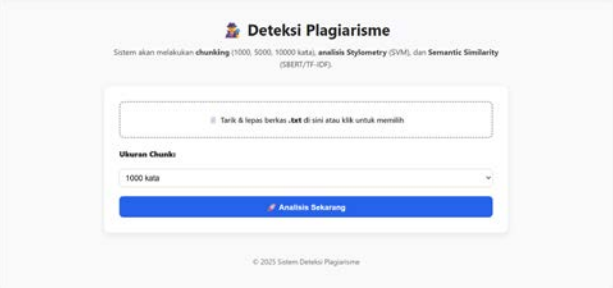
Tabel 11.Tabel Perbandingan *Stylometry* vs SBERT

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

Perbandingan menunjukkan bahwa *Stylometry Similarity* menitikberatkan pada ciri khas gaya penulisan seperti panjang kalimat, pilihan kata, dan struktur sintaksis yang stabil, namun kurang mampu menangkap kesamaan makna secara mendalam. Sebaliknya, *Semantic Similarity* berbasis *SBERT* unggul dalam memahami kesamaan makna, terutama pada kasus parafrasa atau penggunaan sinonim, meski lebih sensitif terhadap teks pendek dan kadang menghasilkan nilai kosong (*NaN*). Karena itu, kombinasi keduanya menjadi pendekatan paling efektif: *stylometry* berperan sebagai penyaring gaya tulisan, sedangkan *semantic similarity* menilai kedekatan makna, sehingga bersama-sama meningkatkan akurasi dan ketepatan deteksi plagiarisme.

3.5 Running App

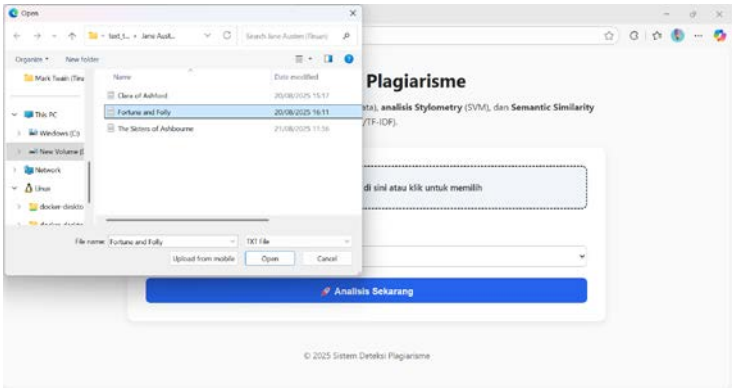
Gambar 1 adalah halaman awal (beranda) sistem deteksi plagiarisme yang berfungsi sebagai titik masuk untuk memulai analisis, di mana pengguna dapat mengunggah file teks (.txt) melalui kotak upload yang tersedia, memilih ukuran chunk (misalnya 1000, 5000, atau 10000 kata) untuk menentukan potongan teks yang akan diproses, lalu menekan tombol “Analisis Sekarang” untuk menjalankan sistem.



Gambar 1. Halaman Utama system deteksi plagiarisme

Setelah teks dipecah, sistem melakukan analisis *stylometry* menggunakan *Support Vector Machine (SVM)* untuk mengenali pola gaya penulisan seperti panjang kalimat, pilihan

kata, dan struktur sintaksis guna mendeteksi perubahan gaya yang mencurigakan sebagai indikasi plagiarisme. Tahap berikutnya adalah analisis kesamaan semantik menggunakan *Sentence-BERT (SBERT)* dan *TF-IDF (Term Frequency–Inverse Document Frequency)* untuk menilai kemiripan makna antar-teks, bahkan jika kalimat telah diparafrase. Kombinasi kedua metode ini memungkinkan sistem mendeteksi plagiarisme berdasarkan gaya penulisan maupun kesamaan makna. Antarmuka sistem dirancang agar pengguna dapat mengunggah file *.txt*, memilih ukuran *chunk*, dan menekan "Analisis Sekarang" untuk memulai proses deteksi secara otomatis.



Gambar 2. Mengupload file text tiruan

Halaman gambar 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 3. Memilih ukuran chunk

Halaman gambar 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.

Hasil analisis pada tiga ukuran *chunk* (1000, 5000, dan 10000 kata) menunjukkan bahwa secara *semantic similarity*, teks tiruan *Fortune and Folly* konsisten meniru makna karya Jane Austen—terutama *Pride and Prejudice* dan *Emma*—dengan skor tertinggi 0.6331 pada *chunk* 1000 kata. Namun, hasil *stylometry* memperlihatkan perbedaan signifikan, di mana gaya penulisan teks tiruan lebih menyerupai Mark Twain pada *chunk* 1000 kata dan Bram Stoker pada ukuran 5000 serta 10000 kata, dengan skor tertinggi 0.7647. Perbedaan ini menunjukkan bahwa meskipun makna dan tema teks tiruan menyerupai Austen, pola linguistik seperti struktur kalimat dan pemilihan kata belum konsisten dengan gaya khasnya. Secara keseluruhan, semakin besar ukuran *chunk*, semakin stabil kesamaan makna yang terdeteksi, namun gaya penulisan tetap menunjukkan pengaruh dari penulis lain.

Tabel 12. 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.

3.6 Interpretasi

Analisis akurasi model *Support Vector Machine (SVM)* menunjukkan kemampuan tinggi dalam mengenali gaya penulisan pada berbagai ukuran *chunk* teks. Hasil evaluasi memperlihatkan akurasi 84.38% pada *chunk* 1000 kata, sedikit menurun menjadi 82.50% pada *chunk* 5000 kata, dan meningkat signifikan hingga 90.48% pada *chunk* 10000 kata. Tren ini mengindikasikan bahwa semakin panjang potongan teks, semakin banyak ciri linguistik yang dapat dimanfaatkan model dalam proses klasifikasi. Meskipun demikian, peningkatan akurasi pada *chunk* panjang perlu diinterpretasikan dengan hati-hati karena jumlah data uji yang lebih sedikit dapat memengaruhi stabilitas hasil. Jane Austen tercatat memiliki *f1-score* tertinggi di atas 0.90 pada semua ukuran teks, menunjukkan konsistensi kuat dalam gaya tulisannya. Mark Twain juga relatif stabil dengan *f1-score* 0.81–0.88, sedangkan Mary Shelley menunjukkan peningkatan performa signifikan pada *chunk*

panjang. Dengan demikian, dapat disimpulkan bahwa semakin panjang teks, semakin kaya ciri *stylometry* yang dapat dikenali oleh model, meskipun terdapat kompromi antara akurasi dan jumlah data uji(Sarwar et al., 2024).

Perbandingan antara *stylometry similarity* dan *semantic similarity (SBERT)* menunjukkan bahwa keduanya memiliki karakteristik analisis yang saling melengkapi. Pendekatan *stylometry* fokus pada pola linguistik dan gaya penulisan khas seorang penulis, seperti panjang kalimat, distribusi tanda baca, serta pilihan kosakata. Metode ini terbukti stabil karena gaya bahasa relatif konsisten dalam satu karya, namun memiliki kelemahan dalam mengenali kesamaan makna yang muncul akibat parafrasa. Sebaliknya, *semantic similarity* berbasis *SBERT* menilai kedekatan makna antar-teks menggunakan representasi embedding semantik, sehingga mampu mendeteksi plagiarisme berbentuk penggantian kata atau struktur kalimat. Hasil penelitian memperlihatkan bahwa Jane Austen memiliki skor *semantic similarity* tertinggi (0.63), menunjukkan teks tiruan mendekati makna aslinya. Namun, Mary Shelley memiliki skor semantik terendah meskipun gaya tulisannya relatif mudah ditiru. Oleh karena itu, kombinasi kedua metode ini penting untuk menghasilkan deteksi plagiarisme yang lebih komprehensif dan akurat(Santander-Cruz et al., 2022).

Interpretasi hasil deteksi plagiarisme menunjukkan hubungan menarik antara kemiripan makna dan gaya penulisan. Berdasarkan analisis *semantic similarity*, teks tiruan *Fortune and Folly* secara konsisten memiliki kedekatan makna dengan karya Jane Austen, terutama *Pride and Prejudice* dan *Emma*. Nilai skor berkisar antara 0.5406 hingga 0.6331, menandakan keberhasilan model dalam meniru struktur ide dan tema utama karya Austen. Namun, hasil analisis *stylometry* justru menunjukkan bahwa gaya penulisan teks tiruan lebih menyerupai Mark Twain pada *chunk* 1000 kata, serta Bram Stoker pada *chunk* 5000 dan 10000 kata. Dengan demikian, kombinasi kedua pendekatan memberikan gambaran yang lebih utuh tentang sejauh mana sebuah teks tiruan meniru karya aslinya dari sisi makna dan gaya bahasa.

4. KESIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa algoritma *Support Vector Machine (SVM)* mampu mengenali gaya penulisan dengan akurasi tinggi, mencapai 90.48% pada *chunk* 10000 kata. Panjang potongan teks berpengaruh signifikan terhadap performa model, di mana teks yang lebih panjang mampu menampilkan ciri linguistik penulis secara lebih jelas, sedangkan teks pendek menghasilkan akurasi yang lebih stabil karena jumlah data uji lebih banyak. Jane Austen terbukti memiliki gaya penulisan paling khas dan konsisten, sementara Mary Shelley menunjukkan peningkatan pengenalan pada teks panjang. Pada aspek deteksi plagiarisme, metode *stylometry similarity* efektif mendeteksi kesesuaian gaya, sedangkan *semantic similarity (SBERT)* unggul dalam menemukan kesamaan makna, sehingga kombinasi keduanya menghasilkan deteksi yang lebih akurat dan menyeluruh. Secara keseluruhan, integrasi analisis berbasis gaya dan makna penting untuk menghadapi bentuk plagiarisme modern yang semakin kompleks, termasuk peniruan gaya dan parafrasa semantik. Saran untuk penelitian selanjutnya adalah memperluas jumlah penulis dan karya agar model lebih *robust*, mengembangkan antarmuka menjadi lebih interaktif melalui aplikasi Android, serta melakukan *deploy* sistem ke layanan web atau *cloud* agar dapat diakses secara daring dan praktis oleh pengguna.

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

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, Issue 1). Springer US. https://doi.org/10.1007/s11042-023-15703-4

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

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

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