ANALISIS FITUR STILOMETRI DAN STRATEGI SEGMENTASI TEKS BERBAHASA INDONESIA PADA SISTEM DETEKSI PLAGIASI INTRINSIK

Skripsi

Diajukan oleh:
SYLVIA PUTRI REJEKI GUNAWAN
71160016

PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
YOGYAKARTA
2020

ANALISIS FITUR STILOMETRI DAN STRATEGI SEGMENTASI TEKS BERBAHASA INDONESIA PADA SISTEM DETEKSI PLAGIASI INTRINSIK

Skripsi

Diajukan kepada Fakultas Teknologi Informasi Program Studi Informatika
Universitas Kristen Duta Wacana
Sebagai salah satu syarat dalam memperoleh gelar Sarjana Komputer

Diajukan oleh:
SYLVIA PUTRI REJEKI GUNAWAN
71160016

PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
YOGYAKARTA
2020

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI SKRIPSI/TESIS/DISERTASI UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademika Universitas Kristen Duta Wacana, saya yang bertanda tangan di bawah ini:

Nama	: Sylvia Putri Rejeki Gunawan
NIM	: 71160016
Program studi	: Informatika
Fakultas	: Fakultas Teknologi Informasi
Jenis Karya	: Skripsi

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Kristen Duta Wacana **Hak Bebas Royalti Noneksklusif** (*None-exclusive Royalty Free Right*) atas karya ilmiah saya yang berjudul:

"Analisis Fitur Stilometri dan Strategi Segmentasi Teks Berbahasa Indonesia pada Sistem Deteksi Plagiasi Intrinsik"

beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti/Noneksklusif ini Universitas Kristen Duta Wacana berhak menyimpan, mengalih media/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat dan mempublikasikan tugas akhir saya selama tetap mencantumkan nama kami sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Yogyakarta
Pada Tanggal : 12 April 2020

Yang menyatakan

(Sylvia Putri Rejeki Gunawan)
NIM.71160016

PERNYATAAN KEASLIAN SKRIPSI

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

ANALISIS FITUR STILOMETRI DAN STRATEGI SEGMENTASI TEKS BERBAHASA INDONESIA PADA SISTEM DETEKSI PLAGIASI INTRINSIK

yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana Komputer pada pendidikan Sarjana Program Studi Informatika Fakultas Teknologi Informasi Universitas Kristen Duta Wacana, bukan merupakan tiruan atau duplikasi dari skripsi kesarjanaan di lingkungan Universitas Kristen Duta Wacana maupun di Perguruan Tinggi atau instansi manapun, kecuali bagian yang sumber informasinya dicantumkan sebagaimana mestinya.

Jika dikemudian hari didapati bahwa hasil skripsi ini adalah hasil plagiasi atau tiruan dari skripsi lain, saya bersedia dikenai sanksi yakni pencabutan gelar kesarjanaan saya.

Yogyakarta, 10 Juli 2020

SYLVIA PUTRI REJEKI GUNAWAN
71160016

HALAMAN PERSETUJUAN

Judul	: Analisis Fitur Stilometri dan Strategi Segmentasi Teks Berbahasa Indonesia pada Sistem Deteksi Plagiasi Intrinsik
Nama	: Sylvia Putri Rejeki Gunawan
NIM	: 71160016
Mata Kuliah	: Skripsi
Kode	: TI0366
Semester	: Genap
Tahun akademik	: 2019/2020

Telah diperiksa dan disetujui
Di Yogyakarta,
Pada Tanggal 9 Juli 2020

Dosen Pembimbing I	Dosen Pembimbing II
Dr. Phil. Lucia Dwi K.,SS., M.A.	Antonius Rachmat C., S.Kom., M.Cs.

HALAMAN PENGESAHAN

ANALISIS FITUR STILOMETRI DAN STRATEGI SEGMENTASI TEKS BERBAHASA INDONESIA PADA SISTEM DETEKSI PLAGIASI INTRINSIK

Oleh: SYLVIA PUTRI REJEKI GUNAWAN / 71160016

Dipertahankan di depan Dewan Penguji Skripsi
Program Studi Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana - Yogyakarta
Dan dinyatakan diterima untuk memenuhi salah satu syarat memperoleh gelar
Sarjana Komputer
pada tanggal 21 Juli 2020

Yogyakarta, 28 Juli 2020
Mengesahkan,

Dewan Penguji:

1. Lucia Dwi Krisnawati, Dr. Phil.
2. Antonius Rachmat C., S.Kom.,M.Cs.
3. Gloria Virginia, S.Kom., MAI, Ph.D.
4.

Dekan

(Restyandito,S.Kom.,MSIS.,Ph.D.)

Ketua Program Studi

(Gloria Virginia,Ph.D.)

UCAPAN TERIMA KASIH

Puji syukur senantiasa penulis panjatkan kepada Tuhan Yang Maha Esa yang telah memberikan kesehatan dan kekuatan dalam penyusunan skripsi ini. Pada kesempatan ini penulis menyampaikan rasa terima kasih yang sebesar-besarnya dan penghargaan yang setinggi-tingginya kepada:

1. Bapak Restyandito, S.Kom., MSIS, Ph.D. selaku Fakultas Teknologi Informasi UKDW.
2. Ibu Gloria Virginia, S.Kom., MAI, Ph.D. selaku Ketua Program Studi Informatika Fakultas Teknologi Informasi UKDW.
3. Ibu Dr. Phil. Lucia Dwi K.,S.S., M.A. selaku dosen pembimbing I, dan
4. Bapak Antonius Rachmat C., S.Kom., M.Cs. selaku dosen pembimbing II yang di sela-sela rutinitasnya namun tetap meluangkan waktunya untuk memberikan petunjuk, dorongan, saran dan arahan sejak rencana penelitian hingga selesainya penulisan skripsi ini.
5. Seluruh dosen di Fakultas Teknologi Informasi, khususnya dosen Program Studi Informatika yang telah memberikan bekal pengetahuan selama penulis menempuh pendidikan di Universitas Kristen Duta Wacana.

Penulis berharap semoga pengorbanan dan segala sesuatunya yang dengan tulus dan ikhlas telah diberikan akan selalu mendapat limpahan rahmat-Nya.

INTISARI

Analisis Fitur Stilometri dan Strategi Segmentasi
Teks Berbahasa Indonesia pada Sistem Deteksi Plagiasi Intrinsik

Plagiasi adalah salah satu jenis pelanggaran hak cipta dengan mencuri kekayaan intelektual. Hadirnya internet memberikan kemudahan bagi siswa untuk melakukan plagiarisme. Bahkan menurut survei, 50% hingga 90% siswa mengaku melakukan plagiarisme setidaknya sekali dalam hidup mereka. Maka dari itu, deteksi plagiasi menjadi penting untuk dilakukan. Sistem deteksi plagiasi intrinsik atau i*ntrinsic plagiarism detection* (*IPD*) berusaha menemukan bagian plagiat dengan mencari segmen teks yang stilometrinya atau gaya penulisannya berbeda dari lainnya. Karena banyaknya jenis fitur stilometri yang ada dan belum adanya ukuran yang pasti mengenai tingkat segmentasi yang paling baik, penulis ingin mencari kombinasi fitur stilometri dan strategi segmentasi yang akurat untuk sistem *IPD* pada teks berbahasa Indonesia. Berdasarkan pengujian, kombinasi yang paling akurat dan praktis adalah segmentasi paragraf dengan fitur stilometri: rata-rata jumlah tanda baca, panjang paragraf, dan rasio *type-token*. Namun, tidak semua jenis teks cocok untuk dideteksi oleh sistem *IPD*. Maka dari itu, sistem *IPD* lebih baik digunakan sebagai rekomendasi bagi penggunanya untuk menemukan perubahan gaya penulisan, bukan untuk menghakimi bagian yang plagiat atau bukan.

Kata kunci: deteksi plagiasi intrinsik, plagiasi, stilometri, segmentasi teks

ABSTRACT

Analysis of Stylometry Features and Text Segmentation Strategies in Intrinsic Plagiarism Detection Systems for Indonesian Language

Plagiarism is a type of copyright infringement by stealing intellectual property. The internet makes it easy for students to practice plagiarism. In fact, 50% to 90% of students admit to had plagiarism. Therefore, plagiarism detection becomes important. Intrinsic plagiarism detection (IPD) systems try to find the plagiarism section by looking for text segments that have different stylometry than others. Because of the many types of stylometry features that exist and there is no definite measure of the best level of segmentation, the author wants to find an accurate combination of stylometric features and segmentation strategies for IPD system in Indonesian language texts. Based on testing, the most accurate combination is paragraph segmentation with stylometric features: average number of punctuation, paragraph length, and type-token ratio. However, not all types of text are suitable for IPD. Therefore, IPD is better used to find writing style change, not to judge the plagiarisms.

Keywords : intrinsic plagiarism detection, plagiarism, stylometry, text segmentation

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB 1
PENDAHULUAN

1.1. Latar Belakang

Plagiasi merupakan suatu isu yang sudah tidak asing lagi mengingat adanya perkembangan teknologi yang memungkinkan penggunanya untuk mendapatkan informasi apapun dari internet. Menurut Maurer, Kappe, dan Zaka (2006), plagiasi adalah pencurian kekayaan intelektual yang telah ada selama manusia menghasilkan karya seni maupun penelitian. Berdasarkan survei yang dilakukan oleh Kuta dan Kitowski (2014), 50% hingga 90% siswa melakukan plagiarisme setidaknya sekali dalam hidup mereka. Ditambah lagi, menurut Krisnawati (2016), hadirnya internet serta kemudahan mengakses berbagai laporan penelitian dan dokumen digital memberikan kemungkinan bagi siswa untuk melakukan plagiarisme seperti yang ditemukan dalam banyak makalah maupun tugas akhir siswa. Plagiasi sendiri termasuk salah satu jenis pelanggaran hak cipta yang telah diatur dalam Undang-Undang Nomor 28 Tahun 2014 tentang Hak Cipta. Maka dari itu, deteksi plagiasi menjadi penting untuk dilakukan.

Plagiasi teks umumnya dideteksi dengan cara membandingkan teks yang diuji dengan beberapa bahan referensi yang diduga merupakan teks sumber. Namun fakta bahwa bahan referensi tidak selalu tersedia maupun jumlah referensi terlalu besar (contohnya jumlah situs web sangat besar di internet) menyebabkan perlu adanya sistem yang dapat mendeteksi plagiasi tanpa perlu membandingkannya dengan teks lain yaitu *intrinsic plagiarism detection* (Halvani, 2015).

Stamatatos (2009) mengungkapkan bahwa *intrinsic plagiarism detection* (*IPD*) atau sistem deteksi plagiasi intrinsik didasarkan pada perubahan atau ketidakkonsistenan gaya penulisan dalam dokumen yang diberikan. Gaya penulisan seseorang dapat dipahami dengan membangun fitur stilometri (*stylometric).* Secara garis besar, *IPD* dilakukan dengan membagi teks menjadi beberapa segmentasi, membangun fitur stilometri di tiap segmentasi, kemudian mendeteksi *outlier* sebagai segmentasi yang diduga berasal dari tindakan plagiasi.

Hingga saat ini, banyak jenis fitur stilometri yang telah dikemukakan oleh para peneliti. Contohnya antara lain n-gram karakter, n-gram kata, rata-rata panjang kata, kalimat, dan paragraf, frekuensi tanda baca, dan lain-lain (Halvani, 2015). Halvani (2015) juga mengatakan bahwa ukuran segmentasi teks yang praktis adalah pertanyaan yang masih terbuka dalam *IPD*. Penelitian yang telah dilakukan sebelumnya juga belum ada yang menggunakan teks berbahasa Indonesia dalam sistem *IPD*. Sehingga dalam penelitian ini, penulis ingin membangun sistem deteksi plagiasi intrinsik pada teks berbahasa Indonesia, serta melakukan analisis kombinasi variasi fitur stilometri dan pengamatan terhadap segmentasi teks. Tujuannya adalah untuk menemukan variasi fitur stilometri serta strategi dekomposisi yang paling akurat dan praktis bagi sistem deteksi plagiasi intrinsik untuk teks berbahasa Indonesia.

1.2. Rumusan Masalah

Dengan latar belakang yang sudah dijelaskan, penulis merumuskan beberapa permasalahan penelitian sebagai berikut :

1. Fitur stilometri apa yang sesuai untuk mendeteksi kasus plagiasi di sistem deteksi plagiasi intrinsik teks berbahasa Indonesia dan bagaimana merepresentasikan fitur-fitur tersebut ke dalam nilai yang tepat dan representatif bagi gaya penulisan seorang pengarang?
2. Segmentasi (dekomposisi) teks dengan tingkat granularitas seperti apakah yang memberikan hasil paling akurat untuk deteksi plagiasi intrinsik teks berbahasa Indonesia?

1.3. Batasan Masalah

Penulis membatasi masalah untuk memudahkan analisis fitur stilometri dan strategi segmentasi teks dalam sistem deteksi plagiasi intrinsik, sebagai berikut:

1. Dokumen yang digunakan adalah dokumen teks karena jenis dokumen yang cocok untuk sistem deteksi plagiasi intrinsik adalah dokumen teks.
2. Teks yang diuji menggunakan teks berbahasa Indonesia. Penulis telah banyak menemukan penelitian mengenai sistem plagiasi intrinsik untuk teks

berbahasa Inggris, namun masih jarang penelitian yang menguji teks berbahasa Indonesia. Maka dari itu, penulis memilih untuk menggunakan teks berbahasa Indonesia dalam membangun sistem deteksi plagiasi intrinsik.

3. Variasi jenis fitur stilometri yang akan digunakan yaitu fitur leksikal berbasis karakter (lihat batasan masalah poin ke-5), fitur leksikal berbasis kata (lihat batasan masalah poin ke-6), fitur sintaksis (lihat batasan masalah poin ke-7), dan fitur struktural (lihat batasan masalah poin ke-8) sesuai dengan jenis fitur stilometri menurut Halvani (2015).
4. Variasi strategi segmentasi/dekomposisi yang digunakan dalam segmentasi teks adalah *structural boundaries*, yaitu per kalimat dan per paragraf. Penulis menggunakan *structural boundaries* karena strategi segmentasi tersebut lebih umum digunakan dan dapat dilakukan dengan cepat tanpa perlu mengidentifikasi elemen khusus dokumen maupun topik tiap bagian dokumen.
5. Fitur leksikal berbasis karakter yang digunakan yaitu frekuensi n-gram karakter untuk segmentasi per kalimat seperti yang dilakukan dalam penelitian Stamatatos (2009), Kuta dan Kitowski (2014), serta Kuznetsov, Motrenko, Kuznetsova, dan Strijov (2016). Nilai n yang digunakan n=3, n=4, dan n=5.
6. Fitur leksikal berbasis kata yang digunakan yaitu frekuensi n-gram kata (untuk segmentasi per paragraf) dengan n=2 dan n=4, panjang kalimat (untuk segmentasi per kalimat), dan rasio *type-token* (untuk segmentasi per paragraf) karena fitur-fitur tersebut lebih umum digunakan dalam *IPD*.
7. Fitur sintaksis yang digunakan yaitu jumlah tanda baca (untuk segmentasi per kalimat) dan rata-rata jumlah tanda baca (untuk segmentasi per paragraf). Untuk fitur sintaksis, penulis memilih menggunakan tanda baca dibandingkan dengan *part-of-speech* karena sampai saat ini belum ada pustaka *part-of-speech tagger* untuk teks berbahasa Indonesia yang akurat.
8. Fitur struktural yang digunakan yaitu panjang paragraf (untuk segmentasi per paragraf) karena fitur ini lebih umum dibanding fitur struktural lainnya

seperti penggunaan tanda tangan, penggunaan kalimat pembuka dan penutup, dan lainnya.

9. Apabila teks yang diunggah ke dalam sistem mengandung kutipan beserta sumber kutipan atau sitasi, maka kutipan dan sumber kutipan tersebut akan tetap diproses dalam komputasi oleh sistem.

1.4. Tujuan Penelitian

Penelitian ini bertujuan untuk membangun sistem deteksi plagiasi intrinsik (*Intrinsic Plagiarism Detection*) untuk teks berbahasa Indonesia. Hal ini dilakukan dengan cara menemukan kombinasi fitur stilometri dan strategi segmentasi teks yang sesuai dan representatif bagi gaya penulisan seseorang. Diharapkan, sistem yang dibangun memberikan hasil yang paling akurat untuk deteksi plagiasi intrinsik teks berbahasa Indonesia.

1.5. Manfaat Penelitian

Beberapa manfaat yang ingin dicapai oleh penulis dengan adanya penelitian ini yaitu:

1. Manfaat penelitian untuk bidang ilmu

   Secara keilmuan, penelitian yang dilakukan akan memberi kontribusi berupa ditemukannya kombinasi fitur stilometri dan strategi segmentasi teks yang cocok serta memberikan hasil yang akurat untuk sistem deteksi plagiasi intrinsik teks berbahasa Indonesia.

2. Manfaat penelitian untuk bidang aplikasi

   Penelitian ini akan menghasilkan aplikasi deteksi plagiasi intrinsik yang dapat digunakan untuk membantu melawan kejahatan pelanggaran hak cipta untuk teks berbahasa Indonesia.

1.6. Metode Penelitian

Metode-metode yang digunakan dalam penelitian ini adalah sebagai berikut:

1. Metode Pengumpulan Data

Pengumpulan data dilakukan secara manual dengan data berupa kumpulan teks berbahasa Indonesia yang sebagian teks-nya merupakan hasil plagiasi. Teks kemudian diberi label pada tiap-tiap bagiannya sebagai segmen plagiasi dan segmen bebas plagiasi supaya hasil sistem *IPD* dapat dievaluasi.

2. Metode Pra-pemrosesan

   Metode pra-pemrosesan yang digunakan yaitu *case folding* dan menghapus karakter-karakter yang tidak relevan dalam penulisan, kemudian dilakukan segmentasi dokumen per paragraf dan per kalimat. Pra-pemrosesan dilakukan menggunakan program yang dibangun oleh penulis.

3. Metode Pembangunan Fitur Stilometri

   Fitur stilometri yang digunakan penulis sebagai representasi gaya penulisan serta yang akan dianalisis untuk menentukan fitur yang paling sesuai dengan deteksi plagiasi intrinsik teks berbahasa Indonesia yaitu:

   a. Frekuensi n-gram karakter

      Untuk pembangunan frekuensi n-gram karakter, penulis menggunakan n=3, n=4, dan n=5 dengan tingkat segmentasi per kalimat. Penulis menghitung frekuensi tiap n-gram karakter per kalimat, sehingga tiap kalimat akan memiliki data array yang merepresentasikan gaya penulisan. Penulis kemudian melakukan komputasi terhadap data yang telah dibentuk untuk menentukan kalimat *outlier* berdasarkan frekuensi n-gram karakter tersebut.

   b. Frekuensi n-gram kata

      Penulis menggunakan n=2 dan n=4 dengan tingkat segmentasi per paragraf. Penulis menghitung frekuensi tiap n-gram kata per paragraf, sehingga tiap paragraf akan memiliki data array yang merepresentasikan gaya penulisan pada paragraf tersebut. Penulis kemudian melakukan komputasi terhadap data yang telah dibentuk untuk menentukan paragraf *outlier* berdasarkan frekuensi n-gram kata tersebut.

   c. Panjang kalimat

Perhitungan panjang kalimat digunakan ketika menggunakan tingkat segmentasi per kalimat. Penulis menghitung jumlah kata pada setiap kalimat, sehingga tiap-tiap kalimat memiliki nilai tunggal berupa panjang kalimat yang digunakan sebagai representasi gaya penulisan.

d. Rasio *type-token*

Rasio *type-token* merupakan perhitungan rasio tipe kata terhadap total kata untuk mengetahui tingkat kekayaan penggunaan kata-kata, variasi, dan redundansi yang penulis gunakan pada segmentasi per paragraf. Penulis menghitung jumlah tipe kata (*type*) yaitu jumlah kata yang berbeda pada paragraf dan membaginya dengan jumlah semua kata yang ada pada paragraf. Masing-masing paragraf akan memiliki nilai tunggal sebagai representasi gaya penulisan.

e. Jumlah tanda baca (*punctuation*)

Perhitungan jumlah tanda baca digunakan ketika menggunakan segmentasi per kalimat. Penulis menghitung jumlah tanda baca pada setiap kalimat, sehingga setiap kalimat dalam dokumen memiliki data tunggal yang merepresentasikan gaya penulisan.

f. Rata-rata jumlah tanda baca (*punctuation*)

Fitur rata-rata jumlah tanda baca dibangun dalam segmentasi per paragraf. Penulis menghitung jumlah semua tanda baca tiap kalimat dalam dokumen, kemudian dihitung rata-ratanya untuk masing-masing paragraf.

g. Panjang paragraf

Fitur panjang paragraf dibangun dengan segmentasi per paragraf. Penulis menghitung jumlah kata dibagi jumlah kalimat dalam tiap paragraf sebagai nilai representasi panjang paragraf.

Pembangunan semua fitur terhadap dokumen yang diinputkan dilakukan secara otomatis menggunakan program yang dibangun oleh penulis.

4. Metode Komputasi *Outlier*

Metode yang digunakan untuk menghitung *outlier* yaitu dengan perhitungan fungsi perubahan gaya bahasa dan perhitungan *boxplot outlier*. Langkah awal perhitungan *boxplot outlier* dilakukan dengan mengurutkan data dari yang terkecil hingga terbesar, kemudian menentukan nilai *Q1, Q3,* dan *IQR*. Setelah itu, penulis mencari nilai batas *inner fences* dan *outer fences,* sehingga didapatkan data-data yang merupakan *major* dan *minor outliers.*

5. Metode Pasca-pemrosesan

   Metode pasca-pemrosesan yang digunakan yaitu *heuristic voting*. *Heuristic voting* menentukan suatu bagian dokumen yang merupakan hasil plagiasi apabila jumlah *outlier*-nya yang terbesar dibandingkan dengan jumlah *outlier* bagian dokumen yang lainnya.

6. Metode Evaluasi

   Metode evaluasi yang digunakan adalah *F1 score* dengan menggunakan granularitas di tingkat kalimat maupun paragraf. Selain itu, penulis juga mencari nilai akurasi, *recall*, dan *precision*.

1.7. Sistematika Penulisan

BAB I PENDAHULUAN membahas tentang gambaran garis besar penelitian yang dilakukan. Bab pendahuluan ini mencakup latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, metode penelitian dan sistematika penulisan. Latar belakang menjelaskan mengenai sistem deteksi plagiasi, temuan penelitian sebelumnya, serta rencana penelitian secara umum. Rumusan masalah merupakan permasalahan yang menjadi inti dilakukannya penelitian. Batasan masalah digunakan untuk memperjelas dan membatasi fokus penelitian. Tujuan penelitian adalah tujuan yang ingin dicapai dengan terealisasinya penelitian yang dilakukan. Metode penelitian berisi langkah-langkah penelitian secara umum. Sistematika penulisan menjelaskan isi tiap bab secara singkat.

BAB II TINJAUAN PUSTAKA DAN LANDASAN TEORI mencakup dua bagian utama yaitu tinjauan pustaka dan landasan teori. Bab ini memuat penelitian serta hasil penelitian yang pernah dilakukan oleh peneliti lain sebelumnya. Tinjauan pustaka berisi tentang tinjauan dari berbagai penelitian sebagai pendukung

penelitian yang dilakukan, sedangkan landasan teori berisi tentang deskripsi teori, metode, serta persamaan matematis yang menjadi prinsip utama penelitian .

BAB III PERANCANGAN SISTEM membahas tentang perancangan sistem deteksi plagiarisme intrinsik yang dibangun. Bab ini menjelaskan kebutuhan sistem, data yang digunakan, dan prosedur untuk membangun sistem deteksi plagiasi intrinsik. Bab ini juga mencakup hingga rancangan pengujian terhadap sistem yang kemudian hasilnya akan digunakan untuk dianalisis.

BAB IV IMPLEMENTASI DAN ANALISIS DATA berisi tentang detail implementasi dari perancangan sistem penelitian beserta analisisnya. Sistem yang telah dibangun sesuai dengan perancangan sistem sebelumnya diimplementasikan kepada data yang telah disiapkan. Hasil dari implementasi ini kemudian dianalisis guna menemukan kombinasi fitur stilometri dan strategi segmentasi teks yang sesuai dan representatif bagi gaya penulisan seseorang pada teks berbahasa Indonesia.

BAB V KESIMPULAN DAN SARAN membahas tentang penarikan kesimpulan dari hasil penelitian dan saran-saran yang dapat digunakan pada penelitian yang berkaitan di masa yang akan datang. Saran-saran ini memuat aktifitas atau langkah-langkah kegiatan dalam riset atau metode dan teknik pengembangan yang belum dilakukan namun dirasa akan memperbaiki kinerja sistem jika langkah-langkah tersebut dilaksanakan pada penelitian mendatang.

BAB 5
KESIMPULAN DAN SARAN

5.1. Kesimpulan

Berdasarkan penelitian yang dilakukan oleh penulis dalam menganalisis kombinasi fitur stilometri dan strategi segmentasi, penulis menarik beberapa kesimpulan sebagai berikut:

1. Strategi segmentasi yang memberikan hasil paling akurat untuk deteksi plagiasi intrinsik teks berbahasa Indonesia adalah segmentasi pada tingkat granularitas paragraf dibandingkan tingkat granularitas kalimat. Segmentasi tingkat granularitas paragraf dapat menghasilkan nilai 0,54 untuk *macro F1-score*, sedangkan nilai *macro F1-score* untuk granularitas kalimat yaitu 0,48.
2. Kombinasi fitur stilometri yang sesuai untuk mendeteksi kasus plagiasi di sistem deteksi plagiasi intrinsik teks berbahasa Indonesia adalah rata-rata jumlah tanda baca, panjang paragraf, dan rasio *type-token*. Kombinasi ini dapat menghasilkan nilai 0,59 untuk *macro F1-score*.
3. Tidak semua jenis teks cocok dideteksi oleh sistem *IPD*. Contoh teks yang tidak cocok yaitu teks yang terlalu pendek (sulit bagi sistem untuk menentukan bagian teks yang berbeda dari lainnya), teks yang sudah melewati penyuntingan (proses penyuntingan adakalanya menyamarkan gaya penulisan seseorang yang sangat kuat), serta teks yang lebih dominan bagian plagiat dibanding bagian orisinilnya (dapat membuat hasil deteksi terbalik). Maka dari itu, hasil deteksi sistem *IPD* dengan menggunakan metode *outlier* sebaiknya digunakan sebagai rekomendasi bagi penggunanya dalam menemukan bagian teks yang gaya penulisannya berbeda.

5.2. Saran

Dengan adanya kekurangan pada sistem deteksi plagiasi intrinsik yang dibangun, saran yang nantinya dapat dikembangan pada sistem ini antara lain:

1. Penambahan deteksi sitasi pada pra-pemrosesan. Bagian yang merupakan hasil pengutipan tidak diikutkan dalam komputasi sistem.
2. Menggunakan n-gram kata atau karakter yang sering muncul sebagai fitur stilometri dibanding menggunakan frekuensi n-gram kata atau karakter.

DAFTAR PUSTAKA

Bensalem, I., Rosso, P., & Chikhi, S. (2014, Oktober). Intrinsic Plagiarism Detection using N-gram Classes. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1459–1464. From https://www.aclweb.org/anthology/D14-1153.pdf

Dawson, R. (2011). How Significant Is A Boxplot Outlier? Journal of Statistics Education, 19(2). doi:10.1080/10691898.2011.11889610

Einsohn, A. (2000). The Copyeditor's Handbook. Berkeley: The Regents of the University of California.

Halvani, O. (2015). Register & Genre Seminar: Towards Intrinsic Plagiarism Detection. From citeseerx.ist.psu.edu: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.465.2185&rep=rep1&type=pdf

IEEE. (tanpa tahun). Plagiarism Levels and Corrective Actions, as taken from Section 8.2.4.D of the PSPB Operations Manual. From ieee.org: https://www.ieee.org/content/dam/ieee-org/ieee/web/org/pubs/Level_description.pdf

Indonesia. (2014). Undang-Undang No. 28 Tahun 2014 tentang Hak Cipta. Jakarta: Sekretariat Negara.

Krisnawati, L. D. (2016). Plagiarism Detection for Indonesian Text. München: Ludwig-Maximilians-Universität München. From https://edoc.ub.uni-muenchen.de/19823/

Kuta, M., & Kitowski, J. (2014). Optimisation of Character n-gram Profiles Method for Intrinsic Plagiarism Detection. In International Conference on Artificial Intelligence and Soft Computing (pp. 500-511). Springer. From https://doi.org/10.1007/978-3-319-07176-3_44

Kuznetsov, M., Motrenko, A., Kuznetsova, R., & Strijov, V. (2016). Methods for Intrinsic Plagiarism Detection and Author Diarization. In CLEF (Working Notes) (pp. 912-919). From

https://pdfs.semanticscholar.org/1011/6d82a8438c78877a8a142be47c4ee8662138.pdf

Lipton, Z. C., Elkan, C., & Naryanaswamy, B. (2014). Thresholding Classifiers to Maximize F1 Score. Machine Learning and Knowledge Discovery in Databases, 8725, 225-239. From https://deepai.org/publication/thresholding-classifiers-to-maximize-f1-score

Maurer, H., Kappe, F., & Zaka, B. (2006). Plagiarism - A Survey. Journal of Universal Computer Science, 12(8), 1050-1084. From http://jucs.org/jucs_12_8/plagiarism_a_survey/jucs_12_08_1050_1084_maurer.pdf

Oberreuter, G., L'Huillier, G., Ríos, S. A., & Velásquez, J. D. (2011). Approaches for intrinsic and external plagiarism detection. Proceedings of the PAN, 4(5), 63. From http://ceur-ws.org/Vol-1177/CLEF2011wn-PAN-OberreuterEt2011.pdf

PUEBI. (2016). Pedoman Umum Ejaan Bahasa Indonesia. Jakarta: Badan Pengembangan dan Pembinaan Bahasa Kementerian Pendidikan dan Kebudayaan.

Stamatatos, E. (2009). Intrinsic Plagiarism Detection Using Character n-gram Profiles. threshold, 2, 38-46. From http://ceur-ws.org/Vol-502/paper8.pdf

Stein, B., Lipka, N., & Prettenhofer, P. (2010). Intrinsic plagiarism analysis. Language Resources and Evaluation, 45(1), 63-82. From https://doi.org/10.1007/s10579-010-9115-y