

**PENERAPAN ALGORITMA SIMHASH UNTUK  
MENDETEKSI KEMIRIPAN TEKS PADA BERITA**

Skripsi



Diajukan oleh:

**MAYESTI ANGGELINA**

**71170247**

**PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI  
UNIVERSITAS KRISTEN DUTA WACANA**

2021

**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 : MAYESTI ANGGELINA  
NIM : 71170247  
Program studi : Informatika  
Fakultas : Teknologi Informasi  
Jenis Karya : Skripsi/Tesis/Disertasi (tulis salah satu)

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:

**“PENERAPAN ALGORITMA SIMHASH UNTUK MENDETEKSI  
KEMIRIPAN TEKS PADA BERITA”**

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 Januari 2022

Yang menyatakan



(MAYESTI ANGGELINA)

71170247

**PENERAPAN ALGORITMA SIMHASH UNTUK  
MENDETEKSI KEMIRIPAN TEKS PADA BERITA**

Skripsi



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

Disusun oleh:

**MAYESTI ANGGELINA**

**71170247**

**PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI  
UNIVERSITAS KRISTEN DUTA WACANA**

2021

## **PERNYATAAN KEASLIAN SKRIPSI**

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

### **PENERAPAN ALGORITMA SIMHASH UNTUK MENDETEKSI KEMIRIPAN TEKS BERITA**

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 keserjanaan 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 keserjanaan saya.

Yogyakarta, 11 Januari 2022



**MAYESTI ANGELINA**  
71170247

**DUTA WACANA**


## HALAMAN PERSETUJUAN

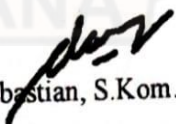
Judul : Penerapan Algoritma Simhash Untuk Mendeteksi  
Kemiripan Teks Pada Berita  
Nama : Mayesti Anggelina  
NIM : 71170247  
Mata Kuliah : Skripsi  
Kode : TI0366  
Semester : Gasal  
Tahun Akademik : 2020/2021

Telah diperiksa dan disetujui  
Di Yogyakarta,  
Pada Tanggal 29 November 2021

Dosen Pembimbing I

Dosen Pembimbing II

  
Dr. Phil. Lucia Dwi K., SS., M.A.

  
Danny Sebastian, S.Kom., M.M., M.T



## HALAMAN PENGESAHAN

### PENERAPAN ALGORITMA SIMHASH UNTUK MENDETEKSI KEMIRIPAN TEKS BERITA

Oleh: MAYESTI ANGGELINA / 71170247


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 14 Desember 2021

Yogyakarta, 11 Januari 2022  
Mengesahkan,

Dewan Penguji:

1. Lucia Dwi Krisnawati, Dr. Phil.
2. Danny Sebastian, S.Kom., M.M., M.T.
3. Willy Sudiarto Raharjo, S.Kom., M.Cs.
4. Joko Purwadi, M.Kom

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

  
Ketua Program Studi  
(Gloria Virginia Ph.D.)

## UCAPAN TERIMAKASIH

Puji syukur penulis Panjatkan Kehadirat Allah Bapa Yang Maha Besar, atas Karunia-Nya. penulis memiliki kesempatan menyelesaikan Laporan Tugas Akhir ini. Pada kesempatan ini penulis ingin menyampaikan wujud rasa terima kasih penulis kepada:

1. Bapak Restyandito, S.Kom.,MSIS.,Ph.D selaku Dekan 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.,SS., M.A. selaku dosen pembimbing I, dan
4. Bapak Danny Sebastian, S.Kom., M.M., M.T selaku dosen pembimbing II yang selalu menyisihkan waktu di sela-sela rutinitasnya yang padat namun tetap meluangkan waktu untuk memberikan arahan, petunjuk sejak awal hingga selesainya penelitian ini.
5. Seluruh dosen di Fakultas Teknologi Informasi, khususnya dosen Program Studi Informatika yang telah memberikan pengetahuan selama masa perkuliahan penulis.
6. Ibu, Bapak, Abang dan Adik tercinta yang telah menjadi motivasi dan tiada henti memberikan dukungan doa untuk saya.
7. Kekasih saya Ari Desrianto Sihombing yang selalu menyemangati, memberikan masukan dan selalu bersedia mendengarkan keluh kesah saya selama proses penelitian ini.

## INTISARI

Penerapan algoritma simhash untuk mendeteksi kemiripan teks pada berita

*Text reuse* bersifat illegal adalah plagiasi, dalam dunia pendidikan plagiasi ini tergolong dalam tindakan yang melanggar peraturan akademik dan Undang-Undang hak cipta yaitu Undang-undang Nomor 28 tahun 2014. Pada beberapa lembaga/organisasi tingkatan *text reuse* bervariasi tergantung toleransi yang telah disepakati. Misalnya di Universitas Kristen Duta Wacana tingkat *text reuse* yang diperbolehkan sekitar 30% diatas itu akan tergolong plagiasi, dan di lembaga/organisasi lain dapat berbeda. Namun, tidak semua *text reuse* dengan tingkat 100% merupakan plagiasi contohnya pada berita. Sebuah *text reuse* dapat dideteksi berdasarkan kemiripan teks tersebut dengan teks yang lain. Maka dari itu penelitian ini akan mengidentifikasi kemiripan teks dengan menerapkan Algoritma Simhash. Algoritma Simhash digunakan untuk mendapatkan fingerprint untuk ekstraksi fitur dalam mendeteksi penggunaan teks kembali (*text reuse*) melalui kemiripan teks antar dokumen. Kemiripan teks dihitung dengan metode *hamming distance*. Berdasarkan hasil pengujian yang dilakukan penulis, mendeteksi teks *duplicate* lebih baik dibandingkan mendeteksi teks *Near-duplicate* karena pada pengujian teks duplikat rata-rata nilai evaluasi recall mencapai 80%. 8 dari 10 dokumen uji sistem bisa menemukan kalimat duplikatnya dengan sempurna. Namun deteksi teks duplikat juga memiliki kekurangan. Terlihat dari nilai rata-rata precision yaitu 27%.

Kata kunci: deteksi kemiripan teks, algoritma *simhash*, *hamming distance*.



## ABSTRACT

The application of simhash algorithm to detect text similarities in news

Text reuse is illegal is plagiarism, in the world of education plagiarism is classified in actions that violate academic regulations and copyright law, namely Law No. 28 of 2014. In some institutions / organizations the level of text reuse varies depending on the tolerance that has been agreed upon. For example, at the Christian University Duta Wacana the level of text reuse allowed about 30% above that will be classified as plagiarism, and in other institutions / organizations can be different. However, not all text reuse with a 100% rate is plagiarism for example in news. A text reuse can be detected based on the similarity of the text to other text. Therefore, this study will identify the similarity of text by applying the Simhash Algorithm. The Simhash algorithm is used to obtain fingerprints for feature extraction in detecting text reuse through text similarities between documents. The similarity of text is calculated by the hamming distance method. Based on the results of the test conducted by the author, detecting *duplicate* text is better than detecting *Near-duplicate* text because in *duplicate* text testing the average recall evaluation value reaches 80%. 8 out of 10 system test documents can find *duplicate* sentences perfectly. But *duplicate* text detection also has drawbacks. Seen from the average value of precision which is 27%.

Keywords: text similarity detection, simhash algorithm, hamming distance.

## DAFTAR ISI

PERNYATAAN KEASLIAN SKRIPSI.....	iii
HALAMAN PERSETUJUAN.....	iv
HALAMAN PENGESAHAN .....	v
UCAPAN TERIMAKASIH.....	vi
INTISARI .....	vii
ABSTRACT.....	viii
DAFTAR ISI.....	ix
DAFTAR TABEL.....	xi
DAFTAR GAMBAR .....	xii
BAB 1 PENDAHULUAN .....	1
1.1. Latar Belakang .....	1
1.2. Rumusan Masalah .....	2
1.3. Batasan Masalah.....	2
1.4. Tujuan Penelitian.....	2
1.5. Manfaat Penelitian.....	2
1.6. Metode Penelitian.....	3
1.7. Sistematik Penulisan .....	4
BAB 2 TINJAUAN PUSTAKA DAN LANDASAN TEORI.....	5
2.1. Tinjauan Pustaka .....	5
2.2. Landasan Teori .....	7
2.2.1. Kemiripan teks .....	7
2.2.2. Pra-pemrosesan .....	7
2.2.3. Fingerprinting .....	8
2.2.4. Algoritma Simhash .....	9
2.2.5. Hamming distance .....	11
2.2.6. Evaluasi.....	13

2.2.7. Copyscape .....	15
<b>BAB 3 PERANCANGAN SISTEM .....</b>	<b>16</b>
3.1.    Kebutuhan Sistem .....	16
3.1.1.    Kebutuhan Fungsional Sistem .....	16
3.1.2.    Kebutuhan Non Fungsional Sistem.....	16
3.2.    Perancangan Sistem.....	17
3.2.1.    Pengumpulan Data .....	17
3.2.2.    Use Case Diagram.....	17
3.2.3.    Blok Diagram Sistem .....	18
3.2.4.    Rancangan Struktur Data .....	20
3.2.5.    Rancangan Pengujian .....	22
3.2.6.    Rancangan Desain Antarmuka.....	25
<b>BAB 4 IMPLEMENTASI DAN ANALISIS SISTEM .....</b>	<b>29</b>
4.1.    Implementasi Tampilan Antarmuka Program .....	29
4.1.1.    Tampilan Antarmuka Menu Home .....	29
4.1.2.    Tampilan Antarmuka Menu About.....	32
4.1.3.    Tampilan Antarmuka Menu How To Use .....	32
4.2.    Implementasi Sistem .....	33
4.2.1.    Pengumpulan Data.....	33
4.2.2.    Pra-pemrosesan.....	34
4.2.3.    Segmentasi teks .....	35
4.2.4.    Pembentukan nilai hash .....	37
4.2.5.    Penyusunan vektor.....	38
4.2.6.    Binerisasi nilai vektor .....	39
4.2.7.    Menghitung kemiripan.....	41
4.2.8.    Memilih data dengan jarak hamming terkecil .....	46
4.3.    Analisis Sistem.....	49
4.3.1.    Hasil pengujian .....	49
4.3.2.    Analisis .....	51

BAB 5 KESIMPULAN DAN SARAN .....	58
5.1. Kesimpulan .....	58
5.2. Saran .....	58
DAFTAR PUSTAKA .....	60
LAMPIRAN.....	62
Lampiran 1. Detail data kalimat hasil deteksi sistem.....	62
Lampiran 2. Source code program .....	62
Lampiran 3. Kartu konsultasi .....	104
Lampiran 4. Lembar Revisi.....	104

## DAFTAR TABEL

Tabel 2.1 <i>Confusion matrix</i> .....	13
Tabel 4.1 Hasil pengujian sistem secara semantik.....	50
Tabel 4.2 Hasil perhitungan nilai evaluasi dan F1 score setiap dokumen terhadap dua tingkat kemiripan.....	50
Tabel 4.3 Data yang benar duplikat dari deteksi sistem terhadap dokumen “Awsdeep_siedoo.txt” .....	52
Tabel 4.4 Detail hasil deteksi sistem tingkat kemiripan <i>Near-duplicate</i> dokumen “Awsdeep_siedoo.txt” .....	53
Tabel 4.5 Evaluasi leksikal hasil deteksi sistem terhadap dokumen “Awsdeep_siedoo.txt” .....	55
Tabel 4.6 Persentase hasil deteksi sistem terhadap dokumen “Awsdeep_siedoo.txt” secara leksikal.....	58
Tabel 4.7 Persentase hasil deteksi sistem terhadap 10 dokumen uji.....	58

## DAFTAR GAMBAR

Gambar 2.1 Contoh proses mendapatkan <i>fingerprint</i> .....	10
Gambar 3.1 Use case diagram.....	18
Gambar 3.2 Blok diagram.....	19
Gambar 3.3 Struktur folder penyimpanan dokumen input.....	21
Gambar 3.4 Contoh penemuan kalimat <i>duplicate</i> .....	23
Gambar 3.5 Contoh kalimat <i>near-duplicate</i> .....	23
Gambar 3.6 Pencatatan hasil deteksi manual.....	24
Gambar 3.7 Pencatatan hasil deteksi sistem.....	25
Gambar 3.8 Desain rancangan antarmuka menu home (tampilan awal).....	26
Gambar 3.9 Desain rancangan antarmuka menu home (tampilan hasil).....	26
Gambar 3.10 Desain rancangan antarmuka menu about.....	27
Gambar 3.11 Desain rancangan antarmuka menu how to use.....	28
Gambar 4.1 Tampilan awal menu home.....	29
Gambar 4.2 Tampilan menu home setelah pengguna memilih dokumen teks yang akan diunggah.....	30
Gambar 4.3 Dokumen yang telah diunggah terdahulu (menjadi dokumen latihan)...	30
Gambar 4.4 Tampilan hasil deteksi.....	31
Gambar 4.5 Dokumen latihan diupdate setelah ada dokumen terunggah.....	31
Gambar 4.6 Tampilan menu about.....	32

Gambar 4.7 Tampilan menu how to use.....	33
Gambar 4.8 Contoh artikel berita yang disalin untuk berita “UKDW Yogyakarta Rayakan Dies Natalis di Tengah Pandemi” .....	34
Gambar 4.9 Teks hasil <i>lower case</i> .....	35
Gambar 4.10 Teks setelah proses pra-pemrosesan.....	35
Gambar 4.11 Teks setelah segmentasi paragraph.....	36
Gambar 4.12 Teks setelah segmentasi kalimat.....	36
Gambar 4.13 Teks hasil menghapus tanda baca.....	37
Gambar 4.14 Contoh token dari sebuah kalimat.....	37
Gambar 4.15 Contoh token hasil penyaringan stopword.....	37
Gambar 4.16 Contoh nilai ASCII dari token sebuah kalimat.....	37
Gambar 4.17 Contoh nilai hash setiap token sebuah kalimat.....	38
Gambar 4.18 Contoh hasil konversi nilai hash ke biner 16 bit.....	38
Gambar 4.19 Contoh <i>fingerprint</i> dokumen.....	40
Gambar 4.20 Contoh isi file Fingerprint.csv.....	40
Gambar 4.21 Contoh isi file Jarak_1.csv.....	42
Gambar 4.22 Contoh isi file Jarak_2.csv.....	43
Gambar 4.23 Contoh isi file Jarak_3.csv.....	44
Gambar 4.24 Contoh isi file Jarak_4.csv.....	45
Gambar 4.25 Contoh data dokumen uji.....	46
Gambar 4.26 Contoh data dokumen latih.....	46



Gambar 4.27 Contoh data yang telah displit.....47

Gambar 4.28 Contoh urutan data pada Jarak\_4.csv.....48

Gambar 4.29 Contoh hasil akhir program.....49

Gambar 4.30 Contoh hasil deteksi sistem pada dokumen “Awsdeep\_siedoo.txt...49



# BAB 1

## PENDAHULUAN

### 1.1. Latar Belakang

Penggunaan teks kembali (*text reuse*) dilakukan untuk menggunakan teks kembali yang telah ada untuk menghasilkan sebuah tulisan yang baru (Clough, Gaizauskas, Piao, & Wilks, 2002). Menurut Krisnawati dan Schulz (2017) *text reuse* terbagi dalam dua jenis yaitu bersifat legal yaitu yang bisa diterima masyarakat yang dan bersifat illegal yaitu yang tidak bisa diterima masyarakat.

*Text reuse* yang bersifat illegal adalah plagiasi, dalam dunia pendidikan plagiasi ini tergolong dalam tindakan yang melanggar peraturan akademik dan Undang-Undang hak cipta yaitu Undang-undang Nomor 28 tahun 2014. Pada beberapa lembaga/organisasi tingkatan *text reuse* bervariasi tergantung toleransi yang telah disepakati. Misalnya di Universitas Kristen Duta Wacana tingkat *text reuse* yang diperbolehkan sekitar 30% diatas itu akan tergolong plagiasi, dan di lembaga/organisasi lain dapat berbeda. Namun, tidak semua *text reuse* dengan tingkat 100% merupakan plagiasi contohnya pada berita. *Text reuse* yang bersifat legal terjadi ketika media berita seperti detik, kompas, metro, dan lain-lain akan menyiarkan sebuah berita yang didapatkan dari kantor berita resmi seperti Kantor Berita Nasional Antara (atau disingkat Perum LKBN Antara) atau institusi lainnya.

Maka dari itu penelitian ini akan mengidentifikasi kemiripan teks dengan dengan menerapkan Algoritma *Simhash*. Algoritma *Simhash* digunakan untuk mendapatkan *fingerprint* yang merupakan ekstraksi fitur dalam mendeteksi penggunaan teks kembali (*text reuse*) melalui kemiripan teks antar dokumen. Program diharapkan mampu menampilkan daftar nama-nama dokumen, id paragraf dokumen yang saling mirip dan mengelompokkan tingkat kemiripan antar paragraf dalam dokumen yaitu *duplicate* atau *near-duplicate*.

## 1.2. Rumusan Masalah

Berdasarkan latar belakang masalah yang terjadi, maka rumusan masalah dalam penelitian ini adalah:

1. Bagaimana menerapkan algoritma *simhash* dalam program deteksi kemiripan teks?
2. Bagaimana efisiensi algoritma *simhash* untuk mendeteksi kemiripan teks?
3. Bagaimana hasil evaluasi program?

## 1.3. Batasan Masalah

Penulis menerapkan beberapa batasan masalah untuk mempermudah penelitian ini yaitu:

1. Objek penelitian adalah berita elektronik berbahasa Indonesia.
2. Berita-berita tersebut dipublikasi 1 tahun terakhir.
3. 30 dokumen latih dan 10 dokumen uji.
4. Segmentasi kalimat.
5. Berita akan dikumpulkan dari situs detik, kompas, tribunnews, liputan6, merdeka, tempo, okezone, suara, JPNN, sindonews, jawapos, suara, dan viva.co.id

## 1.4. Tujuan Penelitian

Tujuan akhir dari penelitian ini adalah membangun program deteksi kemiripan teks (*text reuse*). Deteksi akan dilakukan dengan menerapkan algoritma *simhash* untuk mendapatkan *fingerprint* dari masing-masing kalimat dokumen dan menghitung kemiripan antar *fingerprint* dokumen dengan *hamming distance*. Program yang dibangun akan mampu menampilkan nama-nama dokumen yang saling mirip dan melakukan pengelompokan tingkat kemiripan teks kalimat antar dokumen.

## 1.5. Manfaat Penelitian

Manfaat dari penelitian ini adalah:

1. Program dapat digunakan siapapun dan dimana pun, karena program berbasis website
2. Membantu pengguna mendeteksi kemiripan teks antar file dalam waktu yang singkat.
3. Membantu pengguna untuk melakukan pengelompokan tingkat kemiripan teks antar dokumen.

#### **1.6. Metode Penelitian**

Berikut adalah metode-metode yang penulis gunakan dalam penelitian ini.

1. Pengumpulan Data  
Peneliti akan mengumpulkan data yang diperlukan untuk pengujian program. Data tersebut adalah berita-berita elektronik. Peneliti akan mencari berita yang sama dari beberapa website situs penyiar berita melalui google seperti kompas.com, tribunews.com, detik.com, dan lain-lain.
2. Studi Literatur  
Peneliti akan mempelajari berbagai literatur/artikel/video tentang penerapan algoritma *simhash* pada pemrograman phyton dari berbagai library.
3. Pra-pemrosesan  
Pra-pemrosesan yang digunakan dalam penelitian ini adalah *lower case*, *tokenizing*, *stopword removal* dan menghapus simbol-simbol yang tidak terlalu diperlukan dalam penelitian.
4. Pembangunan Sistem  
Program akan dibangun dengan penerapan algoritma *simhash* pada fitur dokumen dan perhitungan bitwise akan dilakukan dengan *hamming distance*.
5. Metode Evaluasi  
Metode evaluasi yang akan digunakan penulis adalah *F-score*, dan akan mengecek keakuratan program menggunakan *Precision*, dan *Recall*.

## **1.7. Sistematik Penulisan**

BAB I PENDAHULUAN membahas tentang latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, metode penelitian dan sistematik penulisan. Latar belakang membahas tentang program deteksi kemiripan, penyebab terjadinya kemiripan teks, dan dampak kemiripan teks. Rumusan masalah menjelaskan permasalahan yang akan dijawab dengan penelitian ini. Batasan masalah memberikan poin-poin yang menjadi batasan pelaksanaan penelitian ini. Tujuan masalah merupakan tujuan dari penelitian yang dilakukan. Manfaat penelitian berisi manfaat yang akan didapatkan pengguna dari penelitian ini. Metode penelitian berisi metode-metode yang diterapkan dalam penelitian. Sistematikan penulisan menjelaskan isi setiap bab dengan singkat.

BAB II TINJAUAN PUSTAKA DAN LANDASAN TEORI membahas dua hal yaitu tinjauan pustaka dan landasan teori. Tinjauan pustaka berisi ringkasan penelitian-penelitian sebelumnya yang menjadi pendukung pemilihan metode yang diterapkan. Landasan teori berisi penjelasan secara teoritis metode, rumus, dan contoh perhitungan yang diterapkan dalam penelitian.

BAB III PERANCANGAN SISTEM membahas tentang perancangan program deteksi kemiripan teks yang dibangun. Bagian ini menjelaskan kebutuhan program, data latih dan data uji yang akan digunakan untuk program, dan langkah-langkah pembangunan program deteksi kemiripan.

BAB IV IMPLEMENTASI DAN ANALISIS SISTEM membahas tentang detail implementasi dari perancangan sistem penelitian beserta analisisnya. Subbab 4.1 berisi hasil implementasi antarmuka yang dibangun. Subbab 4.2 berisi hasil implementasi sistem pada penelitian ini. Subbab 4.3 berisi analisis pengujian dari penelitian.

BAB V KESIMPULAN DAN SARAN memaparkan tentang penarikan kesimpulan dari hasil analisis pengujian dan saran-saran yang dapat digunakan pada penelitian penelitian yang berkaitan di masa yang akan datang.

## BAB 5

### KESIMPULAN DAN SARAN

#### 5.1. Kesimpulan

Berdasarkan penelitian yang dilakukan oleh penulis dalam menerapkan algoritma simhash untuk mendeteksi kemiripan teks, penulis menarik beberapa kesimpulan berikut:

1. Penerapan algoritma simhash untuk mendeteksi kemiripan dengan tingkat kemiripan dibawah 100% yaitu *Near-duplicate* adalah F1-score bernilai 0.028, precision 0.033 dan recall 0.025.
2. Pengujian algoritma simhash untuk deteksi *Near-duplicate* secara leksikal lebih baik dibandingkan secara semantik.
3. Nilai evaluasi tingkat kemiripan *Duplicate* memiliki rata-rata *precision* 0,27 dan *recall* 0,8 dan F1 *score* 0,37. Nilai *recall* yang cukup tinggi menunjukkan bahwa sistem cukup berhasil mendeteksi kalimat duplikat. Meskipun demikian sistem mengambil terlalu banyak dokumen sehingga membuat nilai *precision* hanya bernilai 0,27.
4. Sistem dirancang untuk menggabungkan segmentasi kalimat yang sangat pendek dengan panjang < 6 kata dengan segmentasi kalimat berikutnya. Sebagai akibatnya kalimat gabungan tersebut tidak terdeteksi sebagai kalimat *duplicate*.

#### 5.2. Saran

Berdasarkan dari pengkajian penelitian ini saran yang dapat dikembangkan pada sistem ini antara lain.

1. Menggunakan n-gram kata atau karakter dalam mendapatkan nilai hash sebelum binerisasi dibanding menggunakan token.



2. Menampilkan hasil luaran sistem dalam bentuk highlight teks yang terduga mirip dibandingkan menampilkan luaran sistem dalam bentuk tabel.



## DAFTAR PUSTAKA

- Schleimer, S., Wilkerson, D. S., & Aiken, A. (2003). *Winnowing: Local Algorithms for Document Fingerprinting*.
- Alamsyah, N. (2017). *Perbandingan Algoritma Winnowing Dengan Algoritma Rabin Karp Untuk Mendeteksi Plagiarisme Pada Kemiripan Teks Judul Skripsi*.
- Alfikri, Z. F., & Purwarianti, A. (2012). *The Construction Of Indonesian-English Cross Language Plagiarism Detection System Using Fingerprinting Technique*.
- Charikar, M. S. (2002). *Similarity Estimation Techniques from Rounding Algorithms*.
- Clough, P., Gaizauskas, R., Piao, S. S., & Wilks, Y. (2002). *METER: MEasuring TExt Reuse*.
- Fitri. (2011, September 23). *Seputar Plagiat dan Autoplgiat*. Retrieved from Kementerian Pendidikan Dan Kebudayaan Lembaga Layanan Pendidikan Tinggi Wilayah XII Maluku Dan Maluku Utara:  
<https://l1dikti12.ristekdikti.go.id/2011/09/23/seputar-plagiat-dan-autoplgiat.html>
- Haryanto, N. C., Krisnawati, L. D., & Chrismanto, A. R. (2020). *Retrieval of source documents in a text reuse system*.
- Krisnawati, L. D. (2016). *Plagiarism Detection for Indonesian Texts*.
- Krisnawati, L. D. (2020). *Detecting Near-Duplicate in Text Reuse*.
- Krisnawati, L. D., & Schulz, K. U. (2013). *Plagiarism Detection for Indonesian Texts*.
- Krisnawati, L. D., & Schulz, K. U. (2017). *Significant Word-based Text Alignment for Text Reuse Detection*.
- Lipton, Z. C., Elkan, C., & Naryanaswamy, B. (2014). *Optimal Thresholding of Classifiers to Maximize F1 Measure*.

- N. Rezaeian, G. N. (2016). Detecting near-duplicates in russian documents through using fingerprint algorithm Simhash.
- Priambodo, J. (2018). Pendeteksian Plagiarisme Menggunakan Algoritma Rabin-Karp Dengan Metode Rolling Hash .
- Shishibori, M., Koizumi, D., & Kita, K. (2011). Fast Retrieval Algorithm for Earth Mover's Distance Using.
- Sood, S., & Loguinov, D. (2011). Probabilistic Near-Duplicate Detection Using Simhash.
- Susanto, D., Basuki, A., & Duanda, P. (2016). Deteksi Plagiat Dokumen Tugas Daring Laporan Praktikum Mata Kuliah Desain Web Menggunakan Metode Naive Bayes.
- Wang, P., Wu, B., Li, X., Wang, L., & Wang, B. (2015). A Simhash-based Generalized Framework for Citation Matching in MapReduce.
- Williams, K., & Giles, C. L. (2013). Near Duplicate Detection in an Academic Digital Library.

