

**OPTIMASI *PRE-PROCESSING* MENGGUNAKAN *REGULAR*
EXPRESSION DAN *COSINE SIMILARITY* UNTUK UJI
KEMIRIPAN KALIMAT BAHASA INDONESIA**

SKRIPSI

Diajukan Untuk Memenuhi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer (S.Kom.)
Pada Progam Studi Teknik Informatika



Oleh :

Ahmad Dzaky Hafidz Musta'in
NPM : 2113020216

**FAKULTAS TEKNIK DAN ILMU KOMPUTER
UNIVERSITAS NUSANTARA PGRI KEDIRI
2025**

Skripsi oleh:

Ahmad Dzaky Hafidz Musta'in
NPM : 2113020216

Judul :

**OPTIMASI *PRE-PROCESSING* MENGGUNAKAN *REGULAR*
EXPRESSION DAN *COSINE SIMILARITY* UNTUK UJI
KEMIRIPAN KALIMAT BAHASA INDONESIA**

Telah Disetujui Untuk Diajukan Kepada Panitia Ujian/Sidang Skripsi
Program Studi Teknik Informatika Fakultas Teknik dan Ilmu Komputer
Universitas Nusantara PGRI Kediri

Tanggal : 14 Juli 2025

Pembimbing I



Ardi Sanjaya, M.Kom.
NIDN. 0706118101

Pembimbing II



Ahmad Bagus Setiawan, ST, M.Kom., MM.
NIDN. 0703018704

Skripsi oleh:

Ahmad Dzaky Hafidz Musta'in
NPM : 2113020216

Judul :

**OPTIMASI *PRE-PROCESSING* MENGGUNAKAN *REGULAR*
EXPRESSION DAN *COSINE SIMILARITY* UNTUK UJI
KEMIRIPAN KALIMAT BAHASA INDONESIA**

Telah dipertahankan di depan Panitia Ujian/Sidang Skripsi
Program Studi Teknik Informatika Fakultas Teknik dan Ilmu Komputer
Universitas Nusantara PGRI Kediri

Pada tanggal : 14 Juli 2025

Dan Dinyatakan telah Memenuhi Syarat

Panitia Penguji :

1. Ketua : Ardi Sanjaya, M.Kom.
2. Penguji I : Resty Wulanningrum, M.Kom
3. Penguji II : Ahmad Bagus Setiawan, ST, M.Kom., MM.



Mengetahui,
Dekan FTIK

Dr. Sulistiono, M.Si
NIDN. 0007076801

HALAMAN PERNYATAAN

Yang bertanda tangan dibawah ini saya,

Nama : Ahmad Dzaky Hafidz Musta'in
Jenis Kelamin : Laki - Laki
Tempat/Tgl Lahir : Kediri, 9 Juli 2003
NPM : 2113020216
Fakultas/Prodi : Teknik dan Ilmu Komputer/Teknik Informatika

Menyatakan dengan sebenarnya, bahwa dalam Skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjaan di suatu perguruan tinggi, dan sepanjang pengetahuan saya tidak dapat karya tulis atau pendapat yang pernah diterbitkan oleh orang lain, kecuali yang secara sengaja dan tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Kediri, 14 Juli 2025

Yang Menyatakan



20
METERAI
TEMPEL
75AMX428331008

Ahmad Dzaky Hafidz Musta'in
NPM : 2113020216

HALAMAN PERSEMBAHAN

Penulisan skripsi ini dengan tulus saya dedikasikan kepada:

1. Ibu saya, yang dengan penuh kesabaran senantiasa mendoakan, memberikan dukungan terbaik, serta menjadi sumber motivasi tak henti-hentinya dalam perjalanan menyelesaikan skripsi ini.
2. Abah saya, terima kasih atas segala didikan dan nilai kehidupan yang telah diberikan. Meski telah tujuh tahun berlalu, kehadiranmu tetap hidup dalam setiap langkah dan pencapaian ini.
3. Kakak saya, yang selalu memberikan semangat sehingga saya dapat menyelesaikan proses ini dengan baik.
4. Keluarga besar, yang telah banyak membantu untuk membiayai kuliah saya. Tanpa dukungan dari mereka, mungkin saya tidak akan bisa sampai di titik ini.
5. Seluruh dosen Universitas Nusantara PGRI Kediri, yang telah memberikan bimbingan, ilmu, dan pelajaran berharga, baik dalam ranah akademik maupun kehidupan sehari-hari.
6. Teman-teman seperjuangan di kampus, yang menjadi tempat berbagi suka dan duka, serta saling menyemangati selama menjalani masa perkuliahan hingga tahap akhir ini.
7. Almamater tercinta, Universitas Nusantara PGRI Kediri, yang telah menjadi tempat saya tumbuh, belajar, dan mengembangkan diri selama masa perkuliahan.
8. Semua pihak yang tidak dapat saya sebutkan satu per satu, yang telah memberikan kontribusi dalam berbagai bentuk dukungan selama proses penyusunan skripsi ini.

Semoga dedikasi ini dapat menjadi penghormatan atas semua dukungan, doa, dan kebersamaan yang telah diberikan.

HALAMAN MOTTO

“Siapapun yang berjuang mencari ilmu karena Allah, akan dijaga setiap langkah perjalanannya sampai ia kembali.”

“Barang siapa yang menempuh jalan untuk mencari ilmu, maka Allah akan mudahkan baginya jalan menuju surga.” – **H.R Muslim**

“Di balik setiap kesulitan dalam belajar, ada kebahagiaan tak terhingga ketika kamu menguasainya.”

“Jangan takut untuk belajar dari kegagalan, karena dalam kegagalan terdapat pembelajaran yang berharga.”

RINGKASAN

Ahmad Dzaky Hafidz Musta'in Optimasi Pre-Processing Menggunakan Regular Expression Dan Cosine Similarity Untuk Uji Kemiripan Kalimat Bahasa Indonesia, Skripsi, Program Studi Teknik Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Nusantara PGRI Kediri, 2025

Kata Kunci : Angka Romawi, *Cosine Similarity*, Kemiripan Kalimat, Optimasi *Pre-Processing*, *Regular Expression*

Pengukuran kemiripan kalimat dalam bahasa Indonesia kerap menghadapi kendala pada format teks yang tidak standar, salah satunya adalah angka Romawi. Pada beberapa sistem, angka sering kali dihapus atau diabaikan selama proses pra-pemrosesan, sehingga menurunkan akurasi hasil pengukuran kemiripan. Penelitian ini mengusulkan solusi dengan mengintegrasikan *regular expression* (regex) dalam tahap *pre-processing* untuk mendeteksi dan mengonversi angka Romawi sebelum dilakukan perhitungan kemiripan menggunakan metode *cosine similarity*. Sistem diuji pada 35 pasangan kalimat yang telah diberi label secara manual. Hasil pengujian menunjukkan bahwa sistem berhasil mengklasifikasikan tingkat kemiripan kalimat dengan akurasi sebesar 89%. Evaluasi dilakukan menggunakan *confusion matrix*, *classification report*, serta analisis terhadap kasus kesalahan klasifikasi. Temuan ini menunjukkan bahwa penambahan penanganan angka Romawi dalam pre-processing dapat meningkatkan keandalan sistem dalam pengolahan teks bahasa Indonesia.

PRAKATA

Puji Syukur dipanjatkan kepada Tuhan Yang Maha Esa, karena atas ridha dan karunianya peneliti dapat menyelesaikan penyusunan laporan penelitian ini. Penulisan ini juga tak lepas dari dukungan pihak yang selalu membantu dalam penulisan penelitian ini. Oleh karenanya peneliti ingin mengucapkan terima kasih kepada:

1. Dr. Zainal Afandi, M.Pd. Selaku Rektor Universitas Nusantara PGRI Kediri.
2. Dr. Sulistiono, M.Si. Selaku Dekan Fakultas Teknik dan Ilmu Komputer Universitas Nusantara PGRI Kediri.
3. Risa Helilintar, M.Kom. Selaku Ketua Program Studi Teknik Informatika Universitas Nusantara PGRI Kediri.
4. Ardi Sanjaya, M.Kom. dan Ahmad Bagus Setiawan, ST, M.Kom., MM. Selaku Dosen Pembimbing Skripsi yang telah dan mengarahkan kami selama mengerjakan skripsi.
5. Kedua Orang Tua saya dan Keluarga atas doa dan dukungannya.
6. Ucapan terimakasih juga disampaikan kepada pihak-pihak lain yang tidak dapat disebutkan satu persatu, yang telah banyak membantu menyelesaikan penulisan penelitian ini.

Disadari penelitian ini masih banyak kekurangan, maka diharapkan kritik dan saran dari berbagai pihak sangat diharapkan. Semoga penelitian ini bermanfaat bagi semua pihak.

Kediri, 14 Juli 2025

Ahmad Dzaky Hafidz Musta'in
NPM. 2113020216

DAFTAR ISI

HALAMAN SAMPUL.....	i
LEMBAR PERSETUJUAN PEMBIMBING SKRIPSI	ii
LEMBAR PENGESAHAN.....	iii
HALAMAN PERNYATAAN	iii
HALAMAN PERSEMBAHAN	v
HALAMAN MOTTO	vi
RINGKASAN.....	vii
PRAKATA.....	viii
DAFTAR ISI	ix
DAFTAR TABEL.....	xi
DAFTAR GAMBAR.....	xii
DAFTAR LAMPIRAN	xiii
BAB I PENDAHULUAN	1
A. Latar Belakang	1
B. Identifikasi Masalah	3
C. Rumusan Masalah	3
D. Batasan Masalah.....	3
E. Tujuan Penelitian	4
F. Manfaat Penelitian	4
BAB II LANDASAN TEORI.....	5
A. Teori dan Penelitian Terdahulu	5
B. Kerangka Berfikir	10
BAB III METODE PENELITIAN	12
A. Desain Penelitian.....	12
B. Instrumen Penelitian	13
C. Jadwal Penelitian.....	16
D. Objek Penelitian/ Subjek Penelitian	16
E. Prosedur Penelitian	18
F. Teknik Analisis Data.....	20

BAB IV HASIL DAN PEMBAHASAN.....	29
A. Hasil Penelitian	29
B. Pembahasan.....	37
BAB V PENUTUP.....	39
DAFTAR PUSTAKA	41
LAMPIRAN	43

DAFTAR TABEL

Tabel 3.1 Perangkat Keras yang Digunakan	14
Tabel 3.2 Perangkat Lunak yang Digunakan.....	14
Tabel 3.3 Jadwal Penelitian	16
Tabel 3.4 Contoh Data	25
Tabel 3.5 Menghitung Frekuensi Kata pada Dua Kalimat	27
Tabel 3.6 Menghitung Frekuensi Kata pada Gabungan Teks	28
Tabel 4.1 Pengujian Fungsional Aplikasi.....	32
Tabel 4.2 Sampel Data Uji (Sebelum Pre-Processing)	34
Tabel 4.3 Sampel Data Uji (Setelah Pre-Processing) dan Hasil Label	35

DAFTAR GAMBAR

Gambar 2.1 Diagram Kerangka Berfikir.....	11
Gambar 3.1 Metode Waterfall	18
Gambar 3.2 <i>Use Case</i> Diagram	20
Gambar 3.3 <i>Activity</i> Diagram	21
Gambar 3.4 Sequence Diagram	22
Gambar 3.5 Desain Struktur Database berbentuk JSON	22
Gambar 3.6 Tampilan Halaman Utama.....	23
Gambar 3.7 Tampilan Cek Kemiripan	23
Gambar 3.9 Tampilan Riwayat	24
Gambar 3.10 Tampilan Tentang Aplikasi.....	24
Gambar 4.1 Halaman Menu Utama	29
Gambar 4.2 Halaman Cek Kemiripan.....	30
Gambar 4.3 Halaman Riwayat Perhitungan	30
Gambar 4.4 Halaman Tentang Aplikasi	31
Gambar 4.5 Evaluasi Visual Confusion Matrix.....	35
Gambar 4.6 Classification Report Evaluasi Sistem.....	36

DAFTAR LAMPIRAN

Lampiran 1 Data Pasangan Kalimat	43
--	----

BAB I

PENDAHULUAN

A. Latar Belakang

Deteksi kemiripan kalimat merupakan topik yang menarik dalam bidang Pemrosesan Bahasa Alami (NLP). Dikutip dari (Sihombing, 2022) pengertian *Natural Language Processing* (NLP) adalah teknik komputasional yang berfungsi untuk memproses dan memahami bahasa alami. Kemampuan untuk mengukur kemiripan antara dua kalimat dapat dimanfaatkan dalam berbagai aplikasi, seperti sistem tanya jawab, pencarian informasi, penerjemahan otomatis, dan *chatbot*. Salah satu metode yang populer untuk mengukur kemiripan antar teks adalah *cosine similarity*, yang menggunakan representasi vektor dari kalimat untuk menghitung derajat kemiripan berdasarkan sudut antara dua vektor tersebut. Metode *cosine similarity* digunakan untuk menentukan seberapa mirip dua dokumen dengan membandingkan sudut antara representasi vektornya (Sugiyamta, 2015). Dalam metode *Cosine Similarity*, dokumen atau teks direpresentasikan dalam bentuk vektor. Proses pencocokan dilakukan dengan membandingkan vektor *term-frequency* dari dua dokumen, yaitu vektor A dan B. Skor kemiripan ini berkisar antara 0 hingga 1 (Sanjaya & Sasongko, 2022).

Namun, terdapat beberapa tantangan dalam mendeteksi kemiripan kalimat, terutama ketika kalimat tersebut mengandung angka. Jika hanya melibatkan angka numerik, proses ini masih dapat diatasi dengan menggunakan fungsi terbilang, yaitu fungsi yang mengubah angka menjadi suatu teks (Sanjaya & Sasongko, 2022). Namun, ketika angka Romawi terlibat, diperlukan penanganan yang lebih kompleks. Angka Romawi memang bisa diproses menggunakan fungsi terbilang, tetapi penting juga untuk memperhatikan konteksnya. Sebagai contoh, angka "V" dan "I" dalam konteks tertentu dibaca sebagai "lima" dan "satu", tetapi bisa juga menjadi huruf yang merupakan bagian dari nama, seperti pada kata "Valentino" dan "Irwan". Oleh karena itu, dalam mendeteksi kemiripan kalimat menggunakan metode seperti *cosine similarity*,

perlu adanya langkah tambahan untuk mengidentifikasi dan memahami konteks penggunaan angka Romawi tersebut. Pada penelitian sebelumnya yang dilakukan oleh (Sanjaya & Sasongko, 2022) yang berjudul “Uji Kemiripan Kalimat Menggunakan Fungsi Terbilang Pada *Pre-Processing* dan *Cosine Similarity* Dalam Bahasa Indonesia” dalam penelitian tersebut untuk menguji kemiripan kalimat menggunakan metode *cosine similarity* dengan proses *pre-processing* berupa fungsi terbilang. Hasilnya menunjukkan bahwa metode ini cukup efektif untuk kalimat yang kompleks, tetapi mengalami penurunan akurasi ketika kalimat tersebut mengandung angka Romawi. Hal ini terlihat dari hasil pengujian tahap kedua yang mengalami sedikit penurunan dari 77,15% menjadi 71,43 %. Tantangan ini muncul karena angka Romawi tidak dikenali oleh fungsi *pre-processing* yang digunakan, sehingga tidak dikonversi ke bentuk numerik yang sesuai untuk diproses. Kemudian peneliti mencoba mengganti angka Romawi tersebut dengan angka numerik dan menjalankan *pre-processing* kembali mendapatkan hasil 85,71 %, yang menegaskan bahwa penanganan angka Romawi merupakan masalah signifikan dalam pengukuran kemiripan kalimat.

Untuk mengatasi masalah ini, dalam penelitian ini, pendekatan baru akan dikembangkan dengan menggunakan *regular expression* (regex) untuk mendeteksi dan mengonversi angka Romawi. *Regular expression* merupakan sebuah bentuk penulisan pola yang digunakan untuk mengenali atau mencocokkan rangkaian karakter tertentu dalam teks (Nur Fadhillah & Rachman, 2023). Dengan menggunakan *regex*, sistem akan mampu mengenali pola angka Romawi dalam kalimat dan mengonversinya ke bentuk numerik atau teks terbilang yang sesuai, sehingga kemiripan kalimat dapat diproses dan mendapatkan hasil yang optimal.

B. Identifikasi Masalah

Merujuk pada uraian yang telah dijelaskan dalam bagian latar belakang, dapat diidentifikasi permasalahan utama dari penelitian ini, yaitu tidak tersedianya fungsi khusus untuk menangani angka Romawi agar kalimat kompleks yang mengandung angka Romawi dapat diproses.

C. Rumusan Masalah

Mengacu pada uraian latar belakang serta identifikasi permasalahan yang telah dijelaskan sebelumnya, maka rumusan masalah dalam penelitian ini dapat dirumuskan sebagai berikut:

1. Bagaimana mengembangkan fungsi *pre-processing* berbasis *regular expression* untuk menangani angka Romawi dalam kalimat bahasa Indonesia?

D. Batasan Masalah

Penelitian ini memiliki sejumlah batasan agar tetap fokus pada ruang lingkup yang telah ditentukan, yaitu:

1. Penelitian ini hanya akan fokus pada kalimat dalam bahasa Indonesia yang mengandung angka Romawi dan numerik.
2. Metode yang digunakan dalam penelitian ini adalah *Cosine Similarity*, tanpa membandingkan dengan metode pengukuran kemiripan lainnya.
3. Pendekatan *pre-processing* yang digunakan terbatas pada penggunaan *regular expression (regex)* untuk deteksi dan konversi angka Romawi ke bentuk numerik atau teks terbilang.
4. Penelitian ini hanya membandingkan pasangan kalimat berdasarkan input yang diberikan oleh pengguna secara langsung, dan tidak mencakup pemrosesan kalimat dari file.
5. Tidak mendalami angka Romawi secara lebih luas, seperti aturan penulisan angka Romawi secara formal atau sejarah penggunaannya, melainkan hanya membahas dalam konteks kemunculannya pada kalimat untuk keperluan pengukuran kemiripan.

6. Penelitian ini tidak mendalami konteks linguistik. Fokus utama adalah pada konversi angka Romawi dalam kalimat untuk mempermudah pengukuran kemiripan.

E. Tujuan Penelitian

Penelitian ini bertujuan untuk:

1. Mengembangkan fungsi *pre-processing* yang mampu mengenali dan mengonversi angka Romawi dalam kalimat bahasa Indonesia menggunakan *regular expression*.

F. Manfaat Penelitian

1. Teoritis
 - a. Memberikan kontribusi dalam pengembangan ilmu di bidang *Natural Language Processing* (Pemrosesan Bahasa Alami), khususnya dalam penanganan angka Romawi pada proses pengukuran kemiripan kalimat.
 - b. Menambah referensi akademis terkait penggunaan metode *cosine similarity* yang dikombinasikan dengan *regular expression (regex)* dalam pengolahan teks.
2. Praktis
 - a. Membantu pengembang sistem berbasis teks, seperti chatbot, sistem tanya jawab, dan aplikasi pencarian informasi, untuk menangani kalimat yang mengandung angka Romawi dengan lebih akurat.
 - b. Menyediakan solusi praktis untuk memproses angka Romawi dalam kalimat menggunakan pendekatan *regex*, sehingga dapat diadopsi dalam pengembangan perangkat lunak lainnya.

DAFTAR PUSTAKA

- Clinton, R. M. R., & Sengkey, R. (2019). Purwarupa Sistem Daftar Pelanggaran Lalulintas Berbasis Mini-Komputer Raspberry Pi. *Jurnal Teknik Elektro dan Komputer*, 8(3), 181–192.
- Dhita, R. L., Faulina, S. T., & Wisnumurti. (2023). Rancang Bangun Aplikasi Layanan Pengaduan Pada Dinas Pendidikan Kabupaten Oku Berbasis Android Menggunakan Android Studio. *Jurnal Informatika dan Komputer*, 14(2), 25–35.
- Fataruba, F. (2018). Penerapan Metode Cosine Similarity Untuk Pengecekan Kemiripan Jawaban Ujian Siswa. *Jurnal Mahasiswa Teknik Informatika*, 2(2), 88–95.
- Guterres, A., Gunawan, & Santoso, J. (2019). Stemming Bahasa Tetun Menggunakan Pendekatan Rule Based. *Teknika*, 8(2), 142–147. <https://doi.org/10.34148/teknika.v8i2.224>
- Kurniadi, D., Haviana, S. F. C., & Novianto, A. (2020). Implementasi Algoritma Cosine Similarity pada sistem arsip dokumen di Universitas Islam Sultan Agung. *Jurnal Transformatika*, 17(2), 124. <https://doi.org/10.26623/transformatika.v17i2.1613>
- Mawanta, I., Gunawan, T. S., & Wanayumini, W. (2021). Uji Kemiripan Kalimat Judul Tugas Akhir dengan Metode Cosine Similarity dan Pembobotan TF-IDF. *Jurnal Media Informatika Budidarma*, 5(2), 726. <https://doi.org/10.30865/mib.v5i2.2935>
- Nur Fadhillah, D., & Rachman, A. (2023). Implementasi Regex Pada Pemberian Komentar Kode Program Html. *Jurnal Advance Research Informatika*, 2(1). <https://www.ejournalwiraraja.com/index.php/JARS>
- Sahrial, R., Fauzi, D. F., & Susilawati, E. (2022). Pemanfaatan JSON Untuk Menampilkan Data Realtime COVID-19 Dengan Model View Presenter. *Jurnal Teknoinfo*, 16(1), 144–149. <https://doi.org/10.33365/jti.v16i1.780>
- Salim, M. A., & Anistyasari, Y. (2017). Pengembangan Aplikasi Penilaian Ujian Essay Berbasis Online Menggunakan Algoritma Nazief Dan Adriani Dengan Metode Cosine Similarity. Dalam *Jurnal IT-EDU* (Vol. 02).
- Sanjaya, A., Bagus Setiawan, A., Mahdiyah, U., Nur Farida, I., & Risky Prasetyo, A. (2023). Pengukuran Kemiripan Makna Menggunakan Cosine Similarity Dan Basis Data Sinonim Kata Measurement Of Meaning Similarity Using Cosine Similarity And Word Synonyms Database. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 10(4). <https://doi.org/10.25126/jtiik.2023106864>

- Sanjaya, A., & Sasongko, S. D. (2022). Uji Kemiripan Kalimat Menggunakan Fungsi Terbilang Pada Pre-Processing Dan Cosine Similarity Dalam Bahasa Indonesia Sentences Similarity Test Using Countable Function On Pre-Processing And Cosine In Indonesian. *Jurnal Ilmiah NERO*, 7(2), 95–104.
- Sibarani, N. S., Munawar, G., & Wisnuadhi, B. (2018). *Analisis Performa Aplikasi Android Pada Bahasa Pemrograman Java dan Kotlin*. 319–324.
- Sihombing, D. O. (2022). Implementasi Natural Language Processing (NLP) dan Algoritma Cosine Similarity dalam Penilaian Ujian Esai Otomatis. *Jurnal Sistem Komputer dan Informatika (JSON)*, 4(2), 396. <https://doi.org/10.30865/json.v4i2.5374>
- Sugiyamta. (2015). Sistem Deteksi Kemiripan Dokumen Dengan Algoritma Cosine Similarity Dan Single Pass Clustering. *Dinamika Informatika*, 7(2), 85–91.
- Toyib, M., Pratama, T. D. K., & Aqil, I. (2024). Penerapan Algoritma CNN Untuk Mendeteksi Tulisan Tangan Angka Romawi dengan Augmentasi Data. *Jurnal Matematika, Ilmu pengetahuan Alam, Kebumihan dan Angkasa*, 2(3), 108–120. <https://doi.org/10.62383/algoritma.v2i3.69>