

**TEXT SUMMARIZATION UNTUK ARTIKEL BAHASA INDONESIA
PADA KORPUS OBJEK BUDAYA DENGAN METODE LEXRANK**

Skripsi



oleh:

**Bernadus Maria Rosario Hardus Tukan
71180280**

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

2022

**TEXT SUMMARIZATION UNTUK ARTIKEL BAHASA INDONESIA
PADA KORPUS OBJEK BUDAYA DENGAN METODE LEXRANK**

Skripsi



Diajukan kepada Program Studi Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Sebagai Salah Satu Syarat dalam Memperoleh Gelar
Sarjana Komputer

Disusun oleh

Bernadus Maria Rosario Hardus Tukan

71180280

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

2022

HALAMAN PENGESAHAN

TEXT SUMMARIZATION UNTUK ARTIKEL BAHASA INDONESIA PADA KORPUS OBJEK BUDAYA DENGAN METODE LEXRANK

Oleh: BERNADUS MARIA ROSARIO HARDUS TUKAN / 71180280

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 25 Oktober 2022

Yogyakarta, 3 November 2022

Mengcsahkan,

Dewan Penguji:

1. Gloria Virginia, S.Kom., MAI, Ph.D.
2. Budi Susanto, SKom., M.T.
3. R. Gunawan Santosa, Drs. M.Si.
4. Yuan Lukito, S.Kom., M.Cs.



Dekan

Ketua Program Studi



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



(Gloria Virginia, Ph.D.)

**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 : Bernadus Maria Rosario Hardus Tukan
NIM : 71180280
Program studi : Informatika
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:

**“TEXT SUMMARIZATION UNTUK ARTIKEL BAHASA INDONESIA
PADA KORPUS OBJEK BUDAYA DENGAN METODE LEXRANK”**

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 : 20 Desember 2022

Yang menyatakan



(Bernadus Maria Rosario Hardus Tukan)
NIM.71180280

PERNYATAAN KEASLIAN SKRIPSI

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

TEXT SUMMARIZATION UNTUK ARTIKEL BAHASA INDONESIA PADA KORPUS OBJEK BUDAYA DENGAN METODE LEXRANK

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, 21 November 2022




Bernadus Maria Rosario Hardus Tukan
71180280

HALAMAN PERSETUJUAN

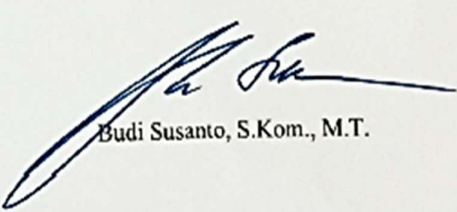
Judul Skripsi : TEXT SUMMARIZATION UNTUK ARTIKEL
BAHASA INDONESIA PADA KORPUS OBJEK
BUDAYA DENGAN METODE LEXRANK
Nama Mahasiswa : BERNADUS MARIA ROSARIO HARDUS TUKAN
NIM : 71180280
Mata Kuliah : Skripsi (Tugas Akhir)
Kode : TI0366
Semester : Genap
Tahun Akademik : 2022/2023

Telah diperiksa dan disetujui di
Yogyakarta,
Pada tanggal 5 Oktober 2022

Dosen Pembimbing I


Gloria Virginia, S.Kom., MAI., Ph.D

Dosen Pembimbing II


Budi Susanto, S.Kom., M.T.

**HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI
TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS \\
SECARA ONLINE UNIVERSITAS KRISTEN DUTA \\
WACANA YOGYAKARTA**

Saya yang bertanda tangan di bawah ini:

NIM : 71180280
Nama : Bernadus Maria Rosario Hardus Tukan
Prodi / Fakultas : Teknologi Informasi / Informatika
Judul Tugas Akhir : TEXT SUMMARIZATION UNTUK ARTIKEL
BAHASA INDONESIA PADA KORPUS
OBJEK BUDAYA DENGAN METODE
LEXRANK

bersedia menyerahkan Tugas Akhir kepada Universitas melalui Perpustakaan untuk keperluan akademis dan memberikan **Hak Bebas Royalti Non Eksklusif** (*Non-exclusive Royalty-free Right*) serta bersedia Tugas Akhirnya dipublikasikan secara online dan dapat diakses secara lengkap (*full access*).

Dengan Hak Bebas Royalti Noneklusif ini Perpustakaan Universitas Kristen Duta Wacana berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk *database*, merawat, dan mempublikasikan Tugas Akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta. Demikian pernyataan ini saya buat dengan sebenar-benarnya.

Yogyakarta, 24 November 2022

Yang menyatakan,



(...71180280 – BERNADUS MARIA ROSARIO HARDUS TUKAN...)

KATA PENGANTAR

Segala puji dan syukur kepada Tuhan yang maha kasih, karena atas segala rahmat, bimbingan, dan bantuan-Nya maka akhirnya Skripsi dengan judul TEXT SUMMARIZATION UNTUK ARTIKEL BAHASA INDONESIA PADA KORPUS OBJEK BUDAYA DENGAN METODE LEXRANK ini telah selesai disusun.

Penulis memperoleh banyak bantuan dari kerja sama baik secara moral maupun spiritual dalam penulisan Skripsi ini, untuk itu tak lupa penulis ucapkan terima kasih yang sebesar-besarnya kepada:

1. Tuhan yang maha kasih,
2. Orang tua yang selama ini telah sabar membimbing dan mendoakan penulis tanpa kenal untuk selama-lamanya.
3. Bapak Restyandito, S.Kom., MSIS., Ph.D. selaku Dekan Fakultas Teknologi Informasi Universitas Kristen Duta Wacana.
4. Ibu Gloria Virginia, S.Kom., MAI, Ph.D. selaku Kepala Program Studi Informatika Fakultas Teknologi Informasi Universitas Kristen Duta Wacana.
5. Ibu Gloria Virginia, S.Kom., MAI, Ph.D. selaku Dosen Pembimbing 1, yang telah memberikan ilmunya dan dengan penuh kesabaran membimbing penulis.
6. Bapak Budi Susanto, S.Kom., M.T. selaku Dosen Pembimbing 2 yang telah memberikan ilmu dan kesabaran dalam membimbing penulis.
7. Keluarga tercinta yang selalu mendukung penulis menyelesaikan skripsi ini.
8. Kristianto Pratama, Gratsia Theodorin, Elza Miyori, Stevani Dwi dan Maria Amanda yang menemani dan menyemangati penulis selama mengerjakan skripsi.
9. Dan juga berbagai pihak lain yang telah mendukung moral, spiritual, dan dana untuk belajar selama ini.

Laporan proposal/skripsi ini tentunya tidak lepas dari segala kekurangan dan kelemahan, untuk itu segala kritikan dan saran yang bersifat membangun guna kesempurnaan skripsi ini sangat diharapkan. Semoga proposal/skripsi ini dapat bermanfaat bagi pembaca semua dan lebih khusus lagi bagi pengembangan ilmu komputer dan teknologi informasi.

Yogyakarta, 24 November 2022
Bernadus Maria Rosario Hardus Tukan



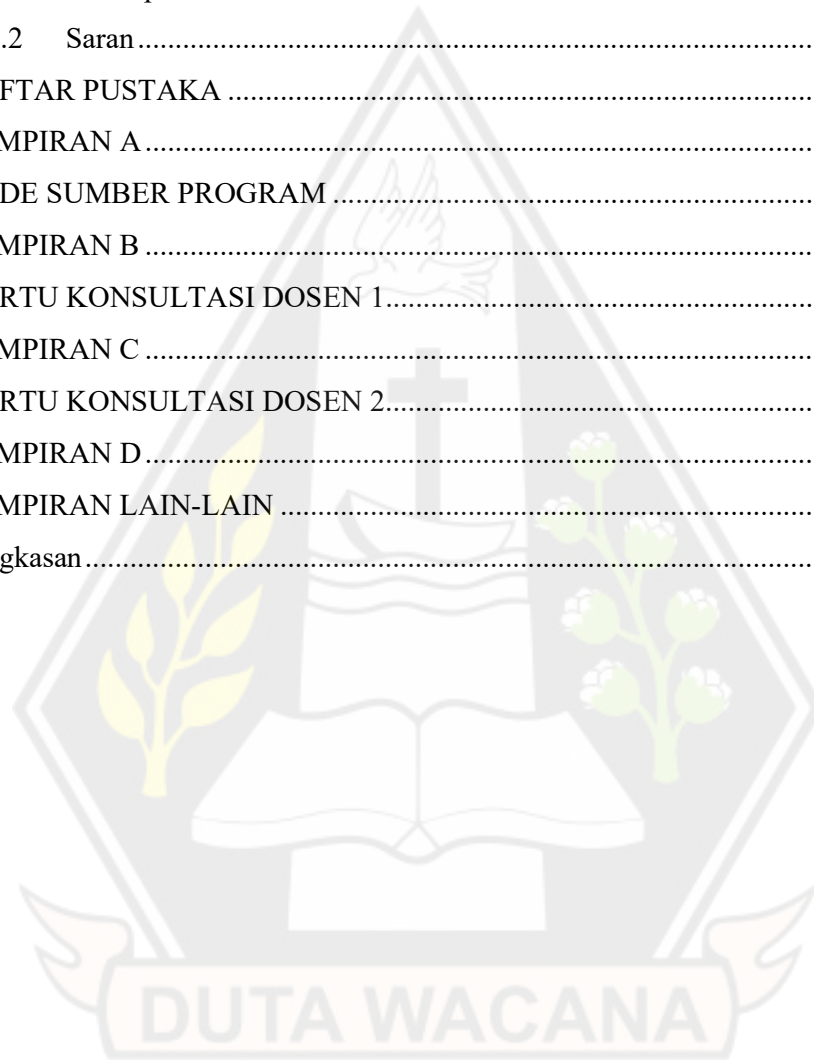
DAFTAR ISI

HALAMAN PENGESAHAN.....	iii
PERNYATAAN KEASLIAN SKRIPSI.....	iv
HALAMAN PERSETUJUAN.....	v
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS \ SECARA ONLINE UNIVERSITAS KRISTEN DUTA \ WACANA YOGYAKARTA	vi
KATA PENGANTAR	vii
DAFTAR ISI.....	ix
DAFTAR TABEL.....	xvi
DAFTAR LAMPIRAN.....	xviii
ABSTRACT.....	xx
BAB I	1
PENDAHULUAN	1
1.1. Latar Belakang Masalah.....	1
1.2. Perumusan Masalah.....	3
1.3. Batasan Masalah.....	3
1.4. Tujuan Penulisan	4
1.5. Manfaat Penulisan	4
1.6. Metodologi Penulisan.....	4
1.7. Sistematika Penulisan.....	6
BAB II.....	7
TINJAUAN PUSTAKA DAN DASAR TEORI	7
2.1 Tinjauan Pustaka	7
2.2 Landasan Teori.....	9
2.2.1 <i>Text Summarization</i>	9
2.2.2 <i>Multi Document Summarization</i>	9
2.2.3 <i>Pre-processing</i>	9
• <i>Case Folding</i>	10

•	<i>Tokenization</i>	10
•	<i>Stopwords Removal</i>	11
•	<i>Lematization</i>	11
•	<i>Labelling Kata dan Phrase Detection</i>	12
2.2.4	<i>LexRank</i>	13
•	<i>Term Frequency – Invers Document Frequency (TF-IDF)</i>	15
•	<i>Power Iteration Method</i>	16
2.2.5	<i>Natural Language Processing</i>	17
2.2.6	<i>ROUGE</i>	18
2.2.7	<i>Kappa Measure</i>	19
BAB III	20
METODOLOGI PENELITIAN	20
3.1	Deskripsi Penelitian.....	20
3.2	Analisis Kebutuhan	20
3.2.1	Spesifikasi Perangkat Pendukung	20
3.2.2	Profile Data	21
3.3	Diagram Alir <i>LexRank</i>	22
3.4	Alur Kerja Program	22
3.5	Tahapan Penelitian	23
3.5.1	Pengambilan Data	23
3.5.2	<i>Unique Token</i>	23
3.5.3	Sumber Data.....	26
3.5.4	<i>Pre-processing</i>	29
3.5.4.1	<i>Original Sentences</i>	29
3.5.4.2	<i>Pre-processing Sentences</i>	30
3.5.4.3	<i>Phrase Sentences</i>	30

3.5.5	<i>Processing</i>	31
3.5.5.1	TF-IDF (<i>Term-Frequency and Invers Document Frequency</i>). ..	32
3.5.5.2	<i>Cosinie Similarity (Idf Modified-Cosine)</i>	32
3.5.5.3	<i>Power Method (Eigen Vector Centrality)</i>	33
3.5.5.4	Ekstraksi 25% Pada <i>Sentences</i>	33
3.6	Profile Responden.....	33
3.7	Tahap Evaluasi.....	34
BAB IV		36
IMPLEMENTASI DAN PEMBAHASAN.....		36
4.1	Antarmuka Aplikasi	36
4.2	<i>Unique Token Processing</i>	37
4.3	Hasil <i>Pre-processing</i>	39
4.3.1	<i>Original Sentences</i>	40
4.3.2	<i>Pre-processing Sentencecs</i>	42
4.3.3	<i>Phrase Sentences</i>	43
4.4	Hasil TF-IDF	43
4.4.1	Hasil TF-IDF <i>LexRank library</i>	44
4.4.2	Hasil TF-IDF dengan perhitungan manual	50
4.5	Hasil <i>Cosine Similarity (Idf Modified-Cosine)</i>	52
4.5.1	Hasil <i>Cosine Similarity (Idf Modified-Cosine) LexRank library</i>	52
4.6	<i>Power Method (Eigen Vector Centrality)</i>	54
4.6.1	Hasil <i>Power Method (Eigen Vector Centrality)</i>	54
4.6.2	Hasil <i>Power Method (Eigen Vector Centrality)</i> pada perhitungan manual	56
4.7	Hasil <i>Kappa Score</i> pada perhitungan manual	58
4.8	Hasil Ringkasan :.....	59
4.9	Evaluasi Hasil Ringkasan.....	65

4.9	Evaluasi <i>Bag of Words LexRank</i>	79
4.10	Perbandingan <i>LexRank</i> dan <i>TextRank</i>	80
BAB V.....		82
KESIMPULAN DAN SARAN.....		82
5.1	Kesimpulan.....	82
5.2	Saran.....	83
DAFTAR PUSTAKA		84
LAMPIRAN A		86
KODE SUMBER PROGRAM		86
LAMPIRAN B		95
KARTU KONSULTASI DOSEN 1.....		95
LAMPIRAN C		96
KARTU KONSULTASI DOSEN 2.....		96
LAMPIRAN D.....		97
LAMPIRAN LAIN-LAIN		97
Ringkasan.....		105

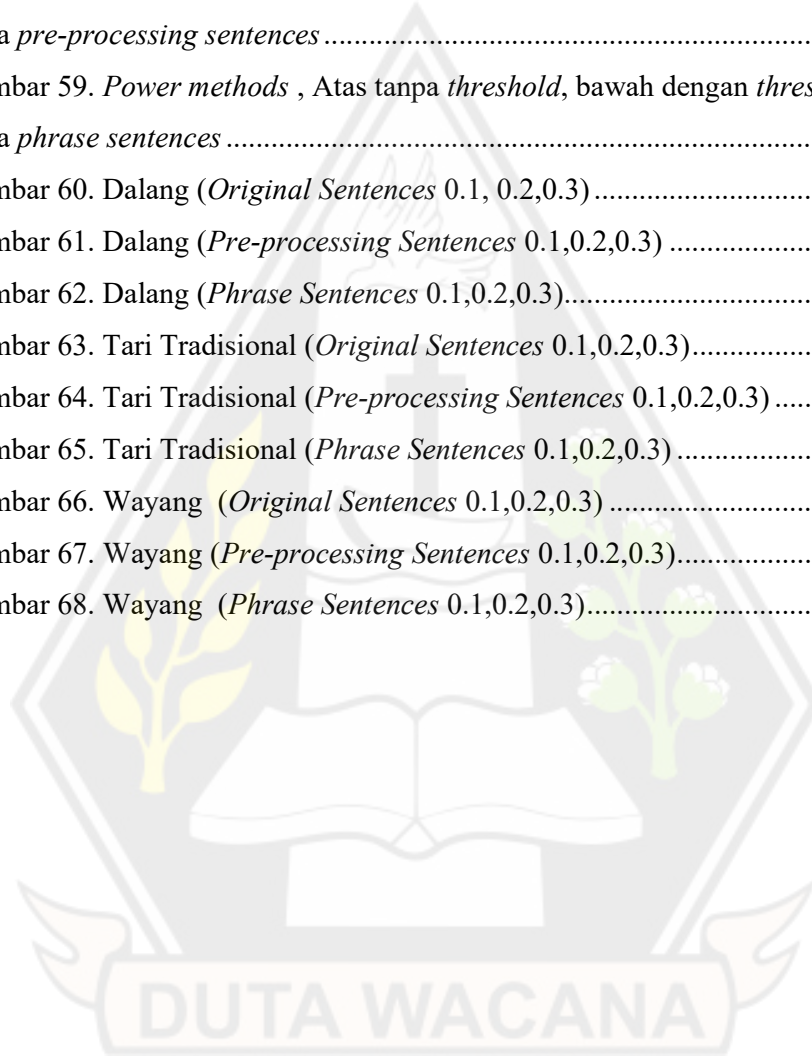


DAFTAR GAMBAR

Gambar 1. Alur Text Summarization.....	9
Gambar 2. <i>Case Folding</i>	10
Gambar 3. <i>Tokenization</i>	11
Gambar 4. <i>Stopwords Removal</i>	11
Gambar 5. <i>Stemming</i>	12
Gambar 6. <i>Connectivity Matrix</i>	14
Gambar 7. Contoh penerapan ambang batas (<i>threshold</i>).....	14
Gambar 8. <i>Broad Classification of NLP</i>	18
Gambar 9. Tampilan Halaman Korpus	21
Gambar 10. Diagram Alir <i>LexRank</i>	22
Gambar 11. Diagram Alur Lexrank Summary.....	23
Gambar 12. Salah satu dokumen yang dilabeling pada Label Studio.....	24
Gambar 13. Labelling Kata dalam bentuk .json.....	26
Gambar 14. Sumber Data untuk <i>pre-processing</i> dan <i>processing LexRank</i>	26
Gambar 15. Kumpulan dokumen yang diinputkan kedalam sistem peringkasan	27
Gambar 16. Kumpulan kalimat hasil <i>pre-processing</i>	27
Gambar 17. Source Code untuk implementasi <i>Library LexRank</i>	28
Gambar 18. Tahapan <i>Pre-processing</i>	29
Gambar 19. Alur <i>Original Sentences</i>	29
Gambar 20. Alur <i>Pre-processing Sentences</i>	30
Gambar 21. Struktur file .json.....	31
Gambar 22. <i>Alur Phrase Sentences</i>	31
Gambar 23. Tahapan Processing.....	32
Gambar 24. Teks Awal	34
Gambar 25. Kalimat Penting.....	35
Gambar 26. Pendapat Responden (khusus form <i>pre-processing sentences</i> dan <i>phrase sentences</i>)	35
Gambar 27. Antarmuka Aplikasi LexSumy.....	36
Gambar 28. <i>Code Unique Token</i> pada program python.....	37
Gambar 29. Kumpulan Tag Kata dari fungsi <i>text_to_string()</i>	38

Gambar 30. Reformat menjadi sekumpulan <i>list</i>	38
Gambar 31. <i>List</i> dari <i>Unique Token</i>	39
Gambar 32. 5 Artikel pada kategori wayang yang masih dalam <i>list of list</i>	40
Gambar 33. 5 Artikel pada kategori wayang yang telah menjadi satu teks	40
Gambar 34. <i>Code Original Sentences</i>	40
Gambar 35. <i>sent_tokenize</i> pada kategori wayang yang dijadikan sebagai <i>original sentences</i>	41
Gambar 36. <i>Code pre-processing sentences</i>	42
Gambar 37. Hasil <i>case folding</i> , <i>tokenization</i> , dan <i>lemmatization</i> menjadi <i>pre-processing sentences</i> pada kategori wayang	42
Gambar 38. <i>Code phrase sentences</i>	43
Gambar 39. <i>Phrase sentences</i> dengan <i>MWETokenizer</i> dan <i>sent_tokenizer</i> pada kategori wayang	43
Gambar 40. <i>Code TF-IDF</i>	44
Gambar 41. <i>Code IDF</i> untuk <i>sm2</i> (Kalimat utama)	44
Gambar 42. Hasil <i>TF IDF</i>	45
Gambar 43. <i>Code TF</i> pada algoritma <i>LexRank</i>	45
Gambar 44. <i>Code idf</i> pada algoritma <i>LexRank</i>	45
Gambar 45. <i>Term Frequency</i> pada <i>library LexRank</i> untuk <i>original sentences</i> ...	46
Gambar 46. <i>Term Frequency</i> menggunakan <i>library LexRank</i> untuk <i>pre-processing sentences</i>	46
Gambar 47. <i>Term Frequency</i> pada <i>library LexRank</i> untuk <i>phrase sentences</i>	47
Gambar 48. Hasil <i>idf</i> pada <i>original sentences</i> menggunakan <i>library LexRank</i>	48
Gambar 49. Hasil <i>idf</i> pada <i>pre-processing sentences</i> menggunakan <i>library LexRank</i>	49
Gambar 50. Hasil <i>idf</i> pada <i>phrase sentences</i> menggunakan <i>library LexRank</i>	50
Gambar 51. Beberapa <i>sentence</i> untuk pengujian <i>idf</i> manual	50
Gambar 52. <i>Code Idf modified-cosine</i> dengan menggunakan function <i>_calculate_similarity_matrix()</i>	52
Gambar 53. <i>Idf Modified-Cosine</i> pada <i>original sentences</i>	53
Gambar 54. <i>Idf Modified-Cosine</i> pada <i>pre-processing sentences</i>	53

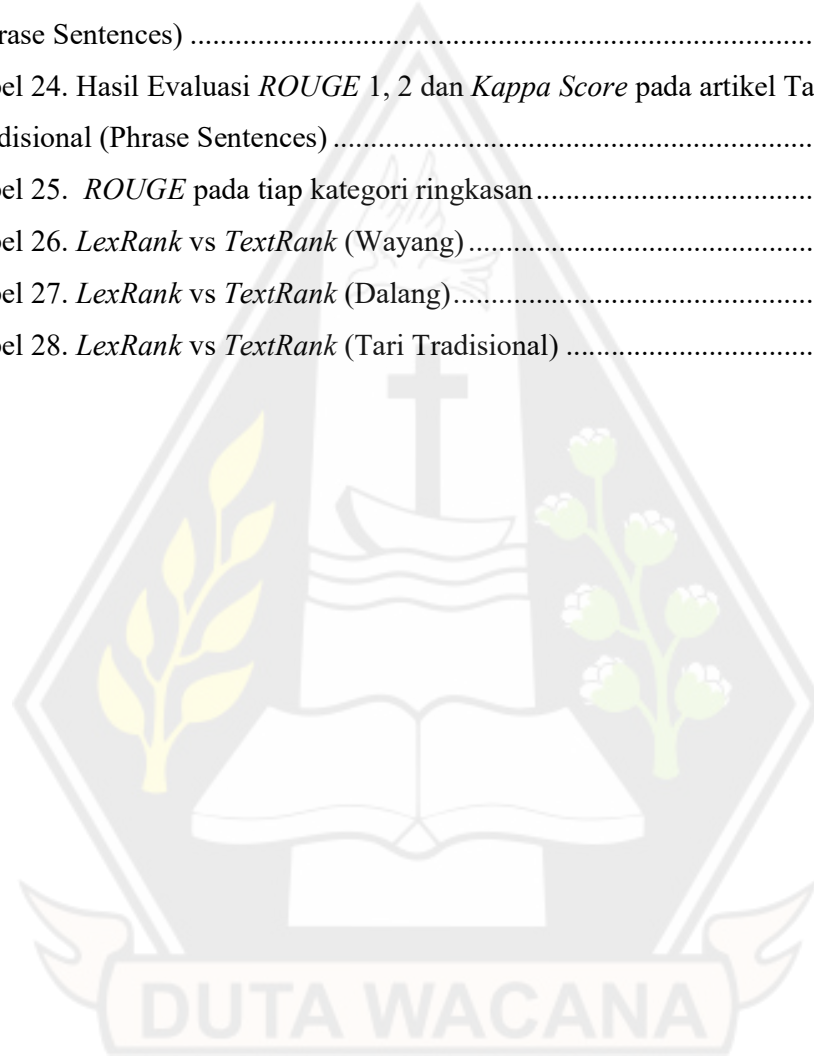
Gambar 55. <i>Idf Modified-Cosine</i> pada <i>phrase sentences</i>	53
Gambar 56. <i>Code power method (eigen vector centrality)</i>	54
Gambar 57. <i>Power methods</i> , Atas tanpa <i>threshold</i> , bawah dengan <i>threshold 0.1</i> pada <i>original sentences</i>	55
Gambar 58. <i>Power methods</i> , Atas tanpa <i>threshold</i> , bawah dengan <i>threshold 0.1</i> pada <i>pre-processing sentences</i>	55
Gambar 59. <i>Power methods</i> , Atas tanpa <i>threshold</i> , bawah dengan <i>threshold 0.1</i> pada <i>phrase sentences</i>	56
Gambar 60. Dalang (<i>Original Sentences 0.1, 0.2,0.3</i>)	60
Gambar 61. Dalang (<i>Pre-processing Sentences 0.1,0.2,0.3</i>)	61
Gambar 62. Dalang (<i>Phrase Sentences 0.1,0.2,0.3</i>).....	61
Gambar 63. Tari Tradisional (<i>Original Sentences 0.1,0.2,0.3</i>).....	62
Gambar 64. Tari Tradisional (<i>Pre-processing Sentences 0.1,0.2,0.3</i>)	62
Gambar 65. Tari Tradisional (<i>Phrase Sentences 0.1,0.2,0.3</i>)	63
Gambar 66. Wayang (<i>Original Sentences 0.1,0.2,0.3</i>)	63
Gambar 67. Wayang (<i>Pre-processing Sentences 0.1,0.2,0.3</i>).....	64
Gambar 68. Wayang (<i>Phrase Sentences 0.1,0.2,0.3</i>).....	64



DAFTAR TABEL

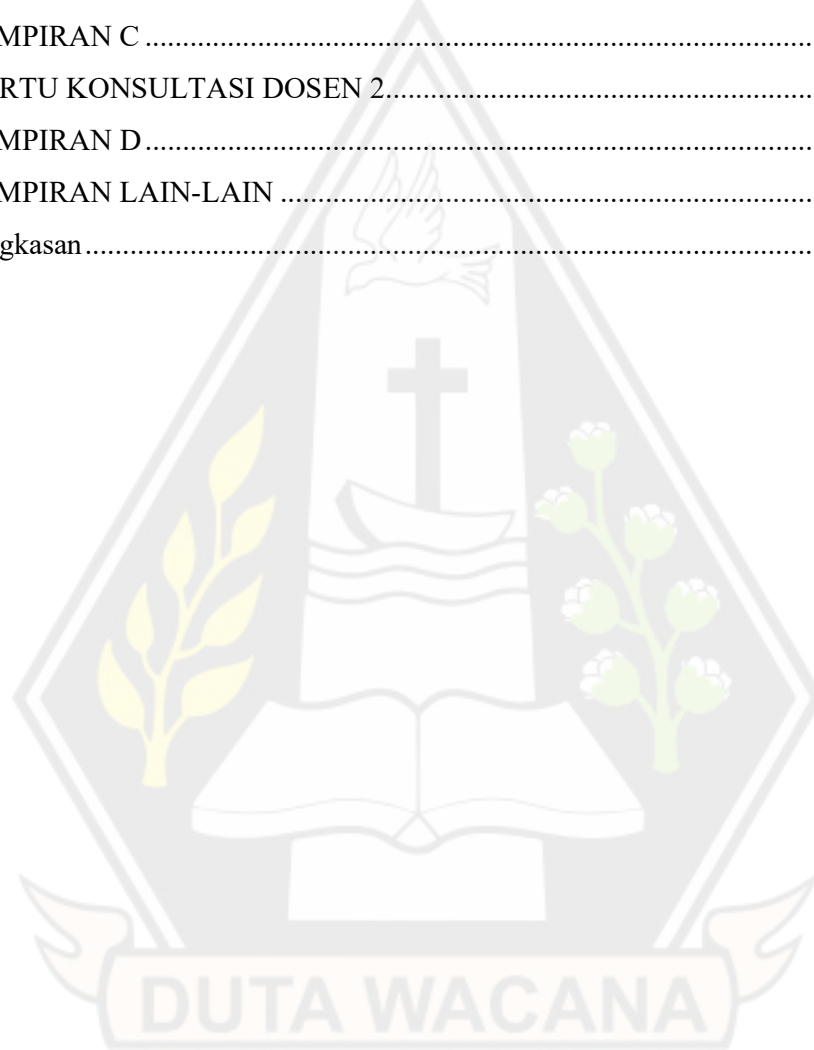
Tabel 1. Tinjauan Pustaka	8
Table 2. Daftar Deteksi frase dan Contoh kata	13
Tabel 3. <i>Strength of agreement</i>	19
Tabel 4. Daftar Label Kata.....	24
Tabel 5. Daftar label kata (lanjutan)	25
Tabel 6. <i>Idf Score</i> setelah dihitung tiap term pada sentence ke-1	51
Tabel 7. Contoh <i>Vector</i>	56
Tabel 8. Iterasi 1.....	57
Tabel 9. Iterasi 2.....	57
Tabel 10. Iterasi 2 (Lanjutan).....	58
Tabel 11. Kappa Matriks.....	58
Tabel 12. Hasil Evaluasi <i>ROUGE</i> 1, 2 dan <i>Kappa Score</i> pada artikel Wayang (Original Sentences).....	67
Tabel 13. Hasil Evaluasi <i>ROUGE</i> 1, 2 dan <i>Kappa Score</i> pada artikel Dalang (Original Sentences).....	68
Tabel 14. Hasil Evaluasi <i>ROUGE</i> 1, 2 dan <i>Kappa Score</i> pada artikel Dalang (Original Sentences) Lanjutan.....	69
Tabel 15. Hasil Evaluasi <i>ROUGE</i> 1, 2 dan <i>Kappa Score</i> pada artikel Tari Tradisional (Original Sentences).....	70
Tabel 16. Hasil Evaluasi <i>ROUGE</i> 1, 2 dan <i>Kappa Score</i> pada artikel Wayang (<i>Pre-processing</i> Sentences).....	71
Tabel 17. Hasil Evaluasi <i>ROUGE</i> 1, 2 dan <i>Kappa Score</i> pada artikel Wayang (<i>Pre-processing</i> Sentences) Lanjutan	72
Tabel 18. Hasil Evaluasi <i>ROUGE</i> 1, 2 dan <i>Kappa Score</i> pada artikel Dalang (<i>Pre-processing</i> Sentences).....	72
Tabel 19. Hasil Evaluasi <i>ROUGE</i> 1, 2 dan <i>Kappa Score</i> pada artikel Dalang (<i>Pre-processing</i> Sentences) Lanjutan.....	73
Tabel 20. Hasil Evaluasi <i>ROUGE</i> 1, 2 dan <i>Kappa Score</i> pada artikel Tari Tradisional (<i>Pre-processing</i> Sentences)	74

Tabel 21. Hasil Evaluasi <i>ROUGE</i> 1, 2 dan <i>Kappa Score</i> pada artikel Wayang (Phrase Sentences)	75
Tabel 22. Hasil Evaluasi <i>ROUGE</i> 1, 2 dan <i>Kappa Score</i> pada artikel Wayang (Phrase Sentences) Lanjutan	76
Tabel 23. Hasil Evaluasi <i>ROUGE</i> 1, 2 dan <i>Kappa Score</i> pada artikel Dalang (Phrase Sentences)	77
Tabel 24. Hasil Evaluasi <i>ROUGE</i> 1, 2 dan <i>Kappa Score</i> pada artikel Tari Tradisional (Phrase Sentences)	78
Tabel 25. <i>ROUGE</i> pada tiap kategori ringkasan	79
Tabel 26. <i>LexRank</i> vs <i>TextRank</i> (Wayang)	80
Tabel 27. <i>LexRank</i> vs <i>TextRank</i> (Dalang)	80
Tabel 28. <i>LexRank</i> vs <i>TextRank</i> (Tari Tradisional)	81



DAFTAR LAMPIRAN

LAMPIRAN A	86
KODE SUMBER PROGRAM	86
LAMPIRAN B	95
KARTU KONSULTASI DOSEN 1.....	95
LAMPIRAN C	96
KARTU KONSULTASI DOSEN 2.....	96
LAMPIRAN D.....	97
LAMPIRAN LAIN-LAIN	97
Ringkasan.....	105



INTISARI

TEXT SUMMARIZATION UNTUK ARTIKEL BAHASA INDONESIA PADA KORPUS OBJEK BUDAYA DENGAN METODE LEXRANK

Pemahaman akan pentingnya informasi dari sebuah laporan atau artikel semakin dibutuhkan di era teknologi saat ini dan orang cenderung tidak menghabiskan waktu untuk mencari substansi dari artikel yang mereka baca. Memberikan sistem peringkasan otomatis menjadi solusi yang bisa ditawarkan ke pembaca.

Pada penelitian ini menggunakan metode *LexRank* (*Lexical PageRank*) untuk melakukan peringkasan teks otomatis dengan teknik *extractive summarization*. Metode *LexRank* menerapkan konsep *Connectivity Matrix* yang memungkinkan suatu kalimat menjadi sangat kuat jika berkoneksi dengan beberapa kalimat lain. *LexRank* juga menggunakan pendekatan berbasis *graph* yaitu memodelkan teks kedalam bentuk *graph* dengan menjadikan teks sebagai *vertex* dan menambahkan *edges* pada *graph* berdasarkan koneksi antar unit teks untuk menentukan tingkat pentingnya kalimat berdasarkan struktur *graph* keseluruhan. Adapun metode-metode terkait untuk memaksimalkan peringkasan *LexRank* yaitu: *TF-IDF*, *Cosine Similarity* dan *Power Iterations (Eigen Vector Centrality)*.

Kalimat yang diekstrak sebesar 25%, dengan ambang batas 0.1, 0.2, 0.3, 0.4, 0.5 dan proses evaluasi menggunakan metode *ROUGE* dengan menghitung *precision*, *recall* dan *f-score*. Selain itu penelitian ini juga melakukan evaluasi untuk menilai performa metode *LexRank* dan metode peringkasan teks lain yaitu *TextRank*.

Kata-kata kunci : Teknologi informasi, peringkasan teks *extractive*, *LexRank*, *ROUGE*, *multi documents*, *Connectivity Matrix*, Pendekatan *Graph*

ABSTRACT

TEXT SUMMARIZATION FOR INDONESIAN ARTICLES ON THE CORPUS OF CULTURAL OBJECTS USING THE LEXRANK METHOD

Understanding the importance of information from a report or article is increasingly needed in today's technology era and people tend not to spend time looking for the substance of the articles they read. Providing an automatic summary system is a solution that can be offered to readers.

In this study, the LexRank (Lexical PageRank) method is used to perform automatic text summarization with extractive summarization techniques. The LexRank method applies the Connectivity Matrix concept which allows a sentence to be very strong if it is connected to several other sentences. LexRank also uses a graph-based approach, namely modeling text into graph form by making text as a vertex and adding edges to the graph based on the connection between text units to determine the level of importance of sentences based on the overall graph structure. While related methods to maximize LexRank summary are: TF- IDF, Cosine Similarity and Power Iterations (Eigen Vector Centrality).

Sentences extracted by 25%, with a threshold of 0.1, 0.2, 0.3, 0.4, 0.5 and the evaluation process using the ROUGE method by calculating precision, recall and f-score. In addition, this study also evaluates the performance of the LexRank method and another text summary method, namely TextRank.

Keywords: Information technology, extractive text summarization, LexRank, ROUGE, multi documents, Connectivity Matrix, Graph Approach

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Kebutuhan mengetahui inti informasi dari sebuah berita atau artikel semakin diperlukan pada era teknologi, masyarakat cenderung kurang suka menghabiskan waktu mereka hanya untuk mencari inti dari artikel yang dibaca. Penggunaan *Text Summarization* merupakan solusi untuk memberikan informasi ringkas ke pembaca tanpa harus menghabiskan banyak waktu. Informasi ringkas yang diberikan tidak semata-mata berupa teks yang dipotong agar terlihat lebih sedikit melainkan informasi ringkasan tersebut sudah diproses sehingga intisari dari artikel dapat tersampaikan ke pembaca. Pada *Text Summarization* terdapat beberapa teknik untuk meringkas informasi salah satunya adalah *extractive summary*. *Extractive summary* berupa penyalinan bagian-bagian teks yang dianggap penting oleh sistem lalu digabungkan menjadi suatu ringkasan. Teks yang disalin berupa kata, kalimat atau paragraf tanpa menambahkan kalimat baru kedalam teks aslinya (Uçkan & Karci, 2020).

Khusus pada Website Korpus Artikel Objek Budaya UKDW “<https://alunalun.info/korpus/>” untuk beberapa waktu terakhir telah mengumpulkan berbagai macam artikel yang berkaitan dengan objek budaya yang ada di Indonesia, Artikel ini disimpan berdasarkan kategori-kategori untuk memudahkan mencari data artikel tertentu. Karena kategori yang telah dikumpulkan belum memiliki ringkasan atau belum menjelaskan intisari dari kategori tersebut maka perlu dilakukan *Text Summarization* sebagai solusi mendapatkan ringkasannya. Untuk mendapatkan intisari bisa diselesaikan dengan cara memilah kata atau kalimat pada setiap artikel lalu secara manual dikumpulkan menjadi satu versi yang lebih singkat untuk mendapatkan ringkasan atau intisari dari kategori artikel yang diinginkan. Umumnya langkah seperti ini bisa dilakukan namun sangat tidak efektif jika kategori artikel memiliki 10 macam artikel dan masing-masing artikelnya mempunyai 5 sampai 8 halaman atau bahkan lebih yang memakan banyak waktu

dan tidak efisien. Memberi *Automatic Text Summarization* dengan menekankan pemilihan kata kunci sehingga membentuk *gold summary* (Klymenko et al., 2020) pada kategori artikel korpus objek budaya UKDW dibutuhkan untuk memberi ringkasan berupa intisari kategori artikel sebelum ditambahkan di katalog repo pada portal web budaya “<https://portal.alunalun.info/>”, tujuannya agar pembaca memahami apa saja yang terdapat pada setiap kategori artikel yang dibaca. Dikarenakan data artikel yang digunakan berbahasa Indonesia, salah satu metode pendukung untuk ringkasan teks adalah metode *LexRank* dengan pendekatan berbasis *graph*. Metode *LexRank* membutuhkan algoritma tertentu untuk peringkasan teks otomatis yaitu algoritma *LexRank*. Peringkasan dengan algoritma *LexRank* bertujuan untuk meringkas dokumen yang banyak sekaligus (*multi documents*).

Pada penulisan ini solusi yang ingin dicapai adalah membuat sistem peringkasan teks otomatis yang diterapkan pada “Portal Web Budaya Indonesia”. Tujuan sistem adalah memberi ringkasan pada setiap kategori artikel yang ada di Portal Web Budaya, sehingga memudahkan pengguna dalam memahami intisari dari kategori yang dikunjungi. Penelitian ini merupakan bagian dari Penelitian Terapan Unggulan Perguruan Tinggi (PTUPT) dengan dana hibah RISTEK-BRIIN berjudul “Portal Objek Budaya Berbasis Semantic Web”.

Berdasarkan latar belakang yang sudah dijelaskan diatas, penelitian ini akan mengembangkan sistem peringkasan otomatis berbasis *multi documents* dengan teknik *extractive* terhadap kategori artikel Objek Budaya .

1.2. Perumusan Masalah

Kategori artikel pada objek budaya yang belum di-intisari menjadi penyebab utama belum bisa dimasukkan kedalam katalog repo pada *project* alun-alun “ <https://app.alunalun.info/>”. Pemberian intisari bisa dilakukan dengan menerapkan *Automatic Text Summarization*. *Automatic Text Summarization* digunakan untuk mempercepat proses ringkasan pada setiap kategori artikel yang banyak dengan bantuan *library LexRank* yang tersedia pada bahasa pemrograman *python* untuk program komputasinya. Berdasarkan masalah yang sudah dijabarkan, penelitian ini akan menerapkan metode *LexRank* dari *library python* untuk melakukan peringkasan teks. Oleh karena dokumen yang diolah adalah berbahasa Indonesia, maka diperlukan juga beberapa *pra-processing* untuk program komputasinya. *Pre-processing* yang digunakan antara lain *case folding*, *tokenization*, *lemmatization*, *labelling* dan *phrase detection* dengan basis kata. Hasil dari penerapan algoritma *LexRank* ini akan dievaluasi menggunakan *ROUGE* untuk menguji kalimat buatan sistem dan kalimat buatan responden.

1.3. Batasan Masalah

Dalam penulisan ini terdapat beberapa batasan yang digunakan yaitu:

- 1) Teks yang digunakan adalah teks yang berbahasa Indonesia.
- 2) Dokumen ringkasan dikhususkan pada kategori artikel objek budaya.
- 3) Sistem hanya akan mengeluarkan output berupa ringkasan untuk setiap kategori artikel.
- 4) Dataset artikel yang diuji berdasarkan data yang didapat dari *Website Korpus Objek Budaya*.
- 5) Kategori yang diringkas ada 3 yaitu Dalang, Tari Tradisional dan Wayang dengan masing-masing kategori sebanyak 20 artikel.
- 6) Hasil evaluasi dengan *ROUGE* menghitung *recall* dan *precision* dari responden tanpa memperhatikan indeks kalimat.
- 7) Pembuatan *Gold Standar* menggunakan *Kappa Score*
- 8) Pemberian Label Kata disesuaikan dengan sumber data

1.4. Tujuan Penulisan

Tujuan penulisan ini adalah membuat sistem peringkasan teks otomatis dengan menerapkan metode *LexRank* yang disediakan oleh *python* untuk memudahkan pengguna untuk mengetahui instisari dari keseluruhan dokumen. Sebagai tools peringkasan teks otomatis untuk portal website budaya Indonesia (<https://portal.alunalun.info/>) dan memberikan sinopsis untuk berbagai korpus budaya yang ada di portal website budaya Indonesia.

1.5. Manfaat Penulisan

Manfaat yang diperoleh dari penelitian ini adalah dengan *Text Summarization* pada korpus artikel objek budaya untuk bahasa Indonesia dapat menyelesaikan masalah berupa abstraksi pada kategori artikel dengan memberi kemudahan bagi pembaca untuk melakukan peringkasan otomatis di portal *web* budaya Indonesia (<https://portal.alunalun.info/>). Sementara pada sisi penulis, mempertajam proses berpikir yang terstruktur dan menambah pengetahuan pada bidang NLP (*Natural Language Processing*).

1.6. Metodologi Penulisan

Beberapa metode yang digunakan untuk mendukung penelitian ini meliputi: Studi literatur, Pengumpulan data, *Pre-processing*, *Processing*, Evaluasi. Tahapan – tahapan penelitian dapat dijabarkan sebagai berikut :

a. Studi Literatur

Studi literatur dilakukan dengan mencari sumber-sumber terkait judul penulisan “*Text Summarization Untuk Artikel Bahasa Indonesia Pada Korpus Objek Budaya Dengan Metode LexRank*”. Studi literatur yang digunakan berupa artikel, jurnal dan skripsi-skripsi terdahulu sebagai referensi dan bahan belajar agar lebih memahami proses dan alur kerja dari Metode *LexRank*.

b. Pengumpulan data

Data yang digunakan berdasarkan kategori untuk setiap dokumen objek budaya yang di input oleh kontributor kedalam website Korpus Artikel Objek Budaya. Kategori yang digunakan yaitu : wayang, tarian tradisional dan dalang.

c. *Pre-processing*

Pre-processing dilakukan pembersihan data terlebih dahulu berupa *case folding, tokenization, lemmatization, labelling* kata dan *phrase detection* sebelum diringkas. *Pre-processing* menggunakan *library NLTK (Natural Language Tool Kit)* dan Sastrawi yang disediakan oleh *python*.

d. *Processing*

Untuk penulisan ini tidak merancang ulang algoritma dari *LexRank*, melainkan menggunakan *library LexRank* yang disediakan *python* untuk proses komputasinya. Penulisan ini lebih menekankan pada proses *Pre-processing* untuk mendapatkan ringkasan yang berkualitas.

e. Evaluasi

Evaluasi yang dilakukan menggunakan *ROUGE* untuk menilai hasil ringkasan secara sistematis dengan menghitung *precision, recall* dan *f-score* dari hasil buatan sistem dan hasil buatan responden. Hasil evaluasi *ROUGE* yang telah didapat dihitung menggunakan *kappa score* untuk mendapatkan *gold standar* berupa nilai valid sebagai nilai akhir evaluasi peringkasan teks. Adapun evaluasi *review* berdasarkan pendapat pribadi dari responden, dimana responden akan melakukan evaluasi pada *google-form* terkait dengan dokumen yang sama, *review* tersebut akan dijadikan sebagai kesimpulan dan saran untuk mengetahui keefektifan sistem yang sudah dibuat.

1.7. Sistematika Penulisan

Sistematika penulisan ini dibagi menjadi 5 bagian utama, dimana setiap bab-nya meliputi :

Bab I Pendahuluan, pada bab ini memaparkan gambaran umum penulisan terkait *Text Summarization* dengan metode *LexRank* yang cakupan sub-bab terdiri dari latar belakang masalah, rumusan masalah, batasan masalah, tujuan penulisan, manfaat penulisan, metodologi penulisan dan sistematika penulisan.

Bab II Tinjauan Pustaka dan Landasan Teori, pada bab ini dijabarkan berbagai jurnal pendukung, definisi yang digunakan, serta formula yang dijadikan acuan untuk penulisan.

Bab III Metodologi Penulisan, pada bab ini berupa cakupan metode yang digunakan pada penelitian ini meliputi: deskripsi penulisan, analisis kebutuhan, diagram alur, tahapan penulisan dan tahapan evaluasi.

Bab IV Implementasi dan pembahasan, pada bab ini meliputi : Antarmuka aplikasi, *unique token processing*, hasil *pre-processing* seperti *original sentences*, *pre-processing sentences* dan *phrase sentences*, *processing Lexrank* : TF-IDF, *Cosine Similarity (idf-modified-cosine)*, *power methods*, *evaluation (ROUGE)* dan *gold standar (kappa score)* .

Bab V Kesimpulan dan Saran, pada bab ini berupa pernyataan singkat dari penulis berdasarkan penjabaran dari hasil penelitian berupa implementasi program peringkasan yang sudah dilakukan dan saran-saran ke depan terkait penelitian serupa.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

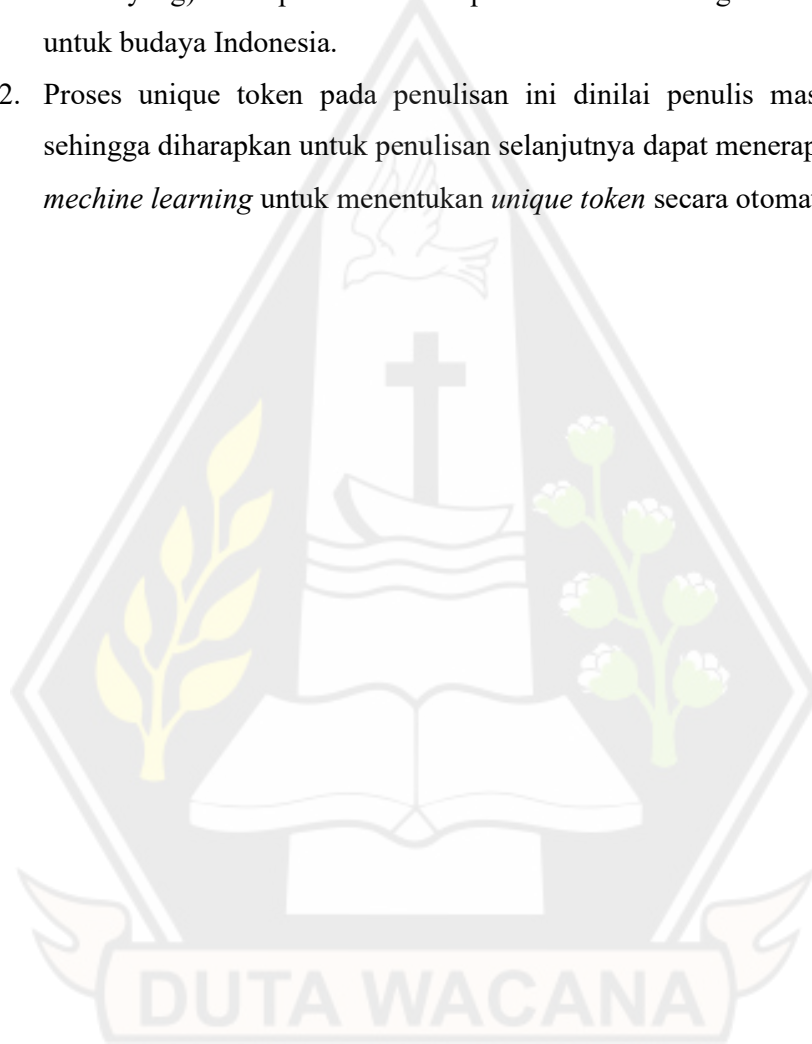
Berdasarkan penelitian yang sudah dilakukan, ada beberapa hal yang bisa disimpulkan diantaranya sebagai berikut:

1. Dari 60 artikel yang diambil dari website korpus dokumen penulis mengambil tiap 5 artikel pada 3 kategori (dalang, tari tradisional dan wayang) untuk proses peringkasan teks. Artikel tersebut telah dievaluasi dan divalidasi sebelum di input ke system peringkasan.
2. Ketiga sentences yang dibuat yaitu *original sentence*, *pre-processing sentence* dan *phrase sentence*, dua diantaranya yaitu *pre-processing sentence* dan *phrase sentence* hanya sebagai fitur “**eksperiman**” dalam penulisan ini dan tidak direkomendasikan sebagai hasil output untuk dibaca user.
3. Evaluasi menggunakan *ROUGE* menunjukkan bahwa, original sentence memiliki hasil *f-score* yang mendekati ringkasan responden dengan nilai range 0.5-0.6 untuk setiap kategori (dalang, tari tradisional dan wayang).
4. Representasi berbasis grafik pada metode *LexRank* dan *TextRank* untuk peringkasan *single documents* dengan menghitung bobot per-kalimat serta pengurutan ringkasan berdasarkan indeks dinilai penulis sebagai faktor utama hasil evaluasi pembandingan yang tidak terlalu jauh, meliputi *LexRank* dengan *f-score* yang rangenya 0.4 -0.6 untuk *ROUGE – 1* dan *ROUGE – 2*
5. Keragaman hasil evaluasi disebabkan karena ringkasan ideal tidak dari satu responden yang sama namun dari beberapa responden sehingga hasil *ROUGE*-nya berbeda pada tiap artikelnya.

5.2 Saran

Adapun saran-saran yang penulis sampaikan untuk pengembangan penulisan serupa kedepannya yaitu :

1. Karena penulisan ini hanya menggunakan 3 kategori (dalang, tari tradisional dan wayang) dirasa perlu dilakukan penulisan untuk ketegori lain khususnya untuk budaya Indonesia.
2. Proses unique token pada penulisan ini dinilai penulis masih manual sehingga diharapkan untuk penulisan selanjutnya dapat menerapkan proses *mechine learning* untuk menentukan *unique token* secara otomatis.



DAFTAR PUSTAKA

- Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in Text Summarization. *Journal of artificial intelligence research*, 22, 457-479.
- Maysyarah, N. (2019). Peringkasan Informasi Otomatis Berita Kriminal Online Bahasa Indonesia Menggunakan LexRank Algorithm
- Vashisht, A. (n.d.). opengenus.org. Retrieved from LexRank method for Text Summarization: <https://iq.opengenus.org/LexRank-text-summarization/>
- Kumar, Y. J., & Salim, N. (2012). Automatic multi document summarization approaches. In KS Gayathri, Received BE degree in CSE from Madras University in 2001 and ME degree from Anna University, Chennai. She is doing Ph. D. in the area of Reasoning in Smart.
- Uçkan, T., & Karıcı, A. (2020). Extractive multi-document Text Summarization based on graph independent sets. *Egyptian Informatics Journal*, 21(3), 145-157.
- Agastya, I. M. A. (2018). Pengaruh Stemmer Bahasa Indonesia Terhadap Peforma Analisis Sentimen Terjemahan Ulasan Film. *Jurnal Tekno Kompak*, 12(1), 18-23.
- Marsyah, Y. Perbandingan Kinerja Algoritma TextRank dengan Algoritma LexRank pada Peringkasan Dokumen Bahasa Indonesia.
- Widman, M. (2020, September 21). *Cohen's Kappa: what it is, when to use it, how to avoid pitfalls*. Retrieved from Knime: <https://www.knime.com/blog/cohens-kappa-an-overview>
- Oktriwina, A. S. (2021, Feb 02). glints.com. Retrieved from NLP: Kecerdasan Buatan yang Bantu Komputer Pahami Bahasa Manusia: <https://glints.com/id/lowongan/natural-language-processing-adalah/>
- Putranto, H. A., Setyawati, O., & Wijono, W. (2016). Pengaruh Phrase Detection dengan POS-Tagger terhadap Akurasi Klasifikasi Sentimen menggunakan SVM. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI)*, 5(4), 252-259.

Setyaningsih, E. R. (2017). Part of Speech Tagger Untuk Bahasa Indonesia Dengan Menggunakan Modifika

Nugroho, K. S. (2019, Juni 18). *Medium.com*. Retrieved from Dasar Text *Pre-processing* dengan Python: <https://ksnugroho.medium.com/dasar-text-pre-processing-dengan-python-a4fa52608ffe>

Zalwert, M. (2021, May 5). Retrieved from LexRank algorithm explained: a step-by-step tutorial with examples: <https://maciejzalwert.medium.com/lexrank-algorithm-explained-a-step-by-step-tutorial-with-examples-3d3aa0297c57>

Kupiec J, Pedersen J, Chen F. 1995. *A trainable document summarizer*. Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval; 1995 Jul 9-13; Seattle, Amerika Serikat. Seattle (US): ACM. hlm 68-73

Hermawan, L., Ismiati, M. B., Bangau, J., 60, N., & Charitas, M. (2020). Pembelajaran Text Preprocessing berbasis Simulator Untuk Mata Kuliah Information Retrieval. *TRANSFORMATIKA*, 17(2), 188–199.

Klymenko, O., Braun, D., & Matthes, F. (2020). Automatic text summarization: A state-of-the-art review. *ICEIS 2020 - Proceedings of the 22nd International Conference on Enterprise Information Systems, 1*, 648–655. <https://doi.org/10.5220/0009723306480655>

Lin, C.-Y., & Hovy, E. (n.d.). *From Single to Multi-document Summarization: A Prototype System and its Evaluation*.

