JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 12 No. 1, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v12i1.3751

PENERAPAN ALGORITMA TEXTRANK DALAM MERANGKUM TEKS WORD DAN PDF

Agustinus Yovi Siang Adi Setiawan[1*], Edwin Alexander*[2]

[1,2]Universitas Katolik Darma Cendika Surabaya; Jl. Dr. Ir. H. Soekarno No.201, Klampis Ngasem, Kec. Sukolilo, Surabaya, Jawa Timur 60117; Telp. 031-5946482

Riwayat artikel:

Received: 25 November 2022
Accepted: 29 Desember 2023
Published: 1 Januari 2024

Keywords: *Text summarization, TextRank*

Corespondent Email:
edwin.alexander@ukdc.ac.id

Abstrak. Ulasan kali ini membahas tentang penerapan algoritma *TextRank* untuk merangkum teks Word dan PDF. Teknologi saat ini memungkinkan informasi berkembang pesat namun juga menimbulkan permasalahan kurangnya waktu untuk meneliti informasi secara lebih mendalam. Algoritma *TextRank* adalah metode pemrosesan bahasa alami berbasis web yang menggunakan pendekatan tanpa pengawasan dan dapat digunakan untuk menghasilkan ringkasan teks otomatis. Metode pencarian yang digunakan meliputi *text preprocessing* dan penggunaan algoritma *TextRank.* Artikel ini juga menjelaskan cara menggunakan *Google Colab* dan *Google Drive* untuk menjalankan algoritma *TextRank* untuk membuat ringkasan. Langkah-langkah untuk menghubungkan Google Colab ke Google Drive, menginstal perpustakaan Python, memanggil data dari Google Drive dan menggunakan algoritma TextRank juga disebutkan dalam artikel ini.

Abstract. This review discusses the implementation of the TextRank algorithm for summarizing Word and PDF texts. Current technology enables rapid information growth but also raises issues regarding the lack of time to delve deeper into information. The TextRank algorithm is a web-based natural language processing method that employs an unsupervised approach and can be used to generate automatic text summaries. The search method used includes text preprocessing and the utilization of the TextRank algorithm. This article also explains how to use Google Colab and Google Drive to execute the TextRank algorithm for summarization. Steps for connecting Google Colab to Google Drive, installing Python libraries, retrieving data from Google Drive, and using the TextRank algorithm are also outlined in this article.

1. PENDAHULUAN

Pada zaman sekarang teknologi menjadi kebutuhan penting bagi manusia yang tidak dapat dipisahkan dari kehidupan, guna membantu memudahkan segala aktivitas dan kegiatan yang dilakukan. Pertumbuhan teknologi pada bidang komunikasi dan komputasi menyebabkan lonjakan besar dalam bidang produksi dan pertukaran informasi yang terjadi saat ini. Perkembangan informasi yang semakin cepat memberikan dampak signifikan pada berbagai aspek kehidupan manusia, baik dalam bidang sosial, pendidikan, maupun lingkungan. Perkembangan informasi yang cepat di internet juga memberikan dampak pada proses pengambilan keputusan [1]. Kecepatan dalam memperoleh informasi memungkinkan untuk mengambil keputusan secara cepat juga, namun hal tersebut juga dapat mengakibatkan kurangnya waktu untuk menyelidiki informasi lebih mendalam, sehingga mengakibatkan kurang tepatnya informasi yang di dapat [2].

Dari survey yang telah dilakukan oleh Asosiasi Penyelenggara Jaringan Internet Indonesia (APJII) pada tahun tahun 2019 – 2020 ada sebanyak 196,7 juta jiwa pengguna aktif internet dari jumlah total keseluruhan penduduk yang mencapai 266,9 juta jiwa. Dari banyaknya pengguna internet tersebut lebih dari 50% pengguna memanfaatkannya untuk membaca artikel berita. Membaca artikel berita merupakan salah satu cara untuk mendapat informasi dengan cepat, tetapi permasalahan yang sering terjadi adalah waktu yang terbatas untuk membaca keseluruhan teks tanpa kehilangan inti dari teks tersebut [3]. Para pengguna internet juga belum tentu memiliki banyak waktu untuk membaca teks yang panjang dikarenakan adanya kesibukan lain. Salah satu cara untuk mengatasi permasalahan tersebut yaitu dengan dibuatnya sistem peringkas teks otomatis yang dapat mengambil ide pokok dari keseluruhan teks yang akan dibaca.

Ringkasan *(summary)* adalah cara yang cukup efektif dalam menyajikan suatu teks yang panjang ke dalam bentuk yang lebih singkat. Ringkasan teks otomatis sangat diperlukan dalam era big data ini, dimana jumlah data teks selalu meningkat secara signifikan dan tidak terstruktur, sehingga untuk membantu menemukan informasi yang relevan dan cepat aplikasi *text summarization* menjadi solusi yang bisa memberikan peran penting dalam meningkatkan efisiensi waktu, mempercepat proses pengambilan keputusan serta mampu memfasilitasi akses cepat terhadap informasi penting. Aplikasi *text summarization* merupakan pembuatan versi pendek dari suatu teks dengan menggunakan aplikasi atau sistem yang berjalan di komputer. Hasilnya berupa poin-poin penting dari teks aslinya [4].

Ada beberapa algoritma yang dapat digunakan untuk membuat peringkas teks otomatis *(automatic text summarization)* salah satunya adalah algoritma textrank. Algoritma textrank merupakan metode pemrosesan bahasa alami berbasis web yang menggunakan pendekatan *unsupervised* dan menggunakan pemodelan hubungan antara kata atau frasa dalam sebuah dokumen [5]. Dalam penelitian ringkasan multi-dokumen yang dilakukan oleh Tendi Arifin, ditemukan bahwa hasilnya memiliki akurasi sebesar 67,35% , recal sebesar 47,83%, presisi sebesar 35,48%, dan nilai f-measure sebesar 40,47% [6]. Meskipun telah memperoleh nilai f-measure metode RVM masih belum bagus dalam kinerjanya, sehingga diperlukan metode lain untuk meningkatkan akurasi dari ringkasan multi dokumen. Penelitian lain yang dilakukan oleh Yuzar Marsyah menjelaskan bahwa hasil rangkuman algoritma textrank mengandung kalimat yang lebih relevan dengan hasil sebesar 95,56%. Algoritma textrank juga lebih cepat dalam mengeluarkan hasil ringkasan daripada algoritma lexrank, yaitu 0,0294 detik dibandingkan dengan 0,4321 detik [7].

2. TINJAUAN PUSTAKA

2.1. *Text Summarization*

Text Summarization atau perangkum teks merupakan metode yang dapat digunakan untuk merangkum dokumen teks yang panjang menjadi lebih ringkas dan memungkinkan representasi singkat yang dapat mencerminkan isi teks yang lebih luas [8]. Merangkum teks dapat dilakukan dengan dua cara pendekatan, yaitu secara abstraktif dan ekstraktif.

A. Abstraktif
   Abstraktif adalah cara merangkum seluruh teks sehingga ringkasannya memiliki kosakata yang lebih bervariasi, bahkan terkadang ada kata-kata yang sama sekali tidak ada dalam teks aslinya [9]. Pendekatan abstraktif lebih sulit tetapi dapat menghasilkan ringkasan dengan kohesi yang tinggi antar kalimat dan lebih alami karena hasil ringkasannya merupakan hasil preparafrase seluruh isi teks seperti halnya ringkasan yang dibuat oleh manusia.
B. Ekstraktif
   Ekstraktif merupakan cara merangkum teks dengan mengambil kalimat-kalimat yang sudah ada sebagai inti teks tanpa modifikasi. Pendekatan secara ekstraktif cenderung lebih mudah, akan tetapi sering sekali menghasilkan ringkasan dengan kohesi antar kalimat yang rendah [10].

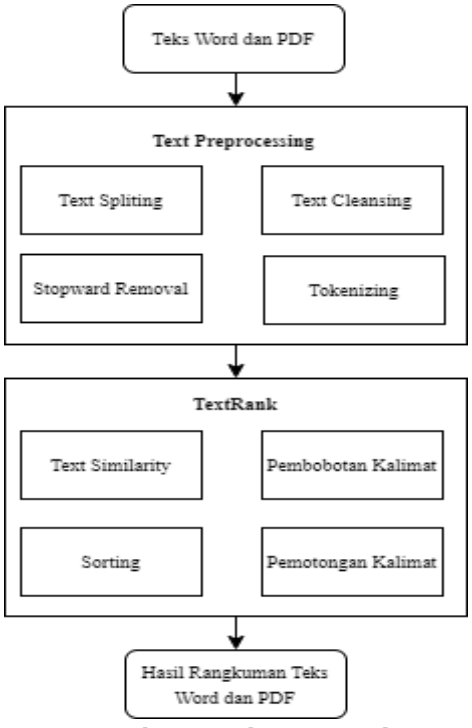
2.2. TextRank

TextRank adalah algoritma yang digunakan untuk mendapatkan kata-kata paling penting dalam sebuah dokumen teks. TextRank berbasis graf untuk memberi peringkat pada teks dan kalimat-kalimat teks di representasikan sebagai simpul atau titik dalam grafik [11]. Peneliti lain juga menjelaskan bahwa metode TextRank merupakan algoritma perangkum berbasis graf yang dibangun berdasarkan metode PageRank yang terdiri atas vertex yang mempresentasikan kalimat pada dokumen dan *edge* yang mempresentasikan hubungan kemiripan antar kalimat [12].

3. METODE PENELITIAN

Untuk memfasilitasi proses penelitian dalam upaya mencapai tujuan dari penelitian ini, maka diperlukan sebuah susunan atau tahapan yang disusun dengan sistematis mulai dari tahap awal hingga tahap akhir sampai mendapat hasil yang diinginkan. Sebelum mendapat hasil akhir tentunya terlebih dahulu melakukan studi literatur terkait penelitian yang dilakukan yaitu mengenai Penerapan

Algoritma TextRang dalam Merangkum Teks Word dan PDF. Tahap penelitian yang dilakukan pada penelitian ini dapat dilihat pada Gambar1.



Gambar 1. Tahapan Penelitian

3.1. Teks Word dan PDF

Teks Word merupakan sebuah teks yang mengacu pada teks atau dokumen yang dibuat atau disimpan menggunakan perangkat lunak Microsoft Word. Dokumen Word dapat mencangkup berbagai jenis dokumen seperti surat, laporan, esai, dan banyak lainnya. Sedangkan Teks PDF merupakan sebuah teks yang disusun dalam format PDF *(Portable Document Format).* PDF merupakan format file yang digunakan untuk presentasi dokumen elektronik dengan tata letak tetap, tidak bergantung perangkat keras, sistem komputer, dan perangkat lunak yang digunakan untuk membuatnya.

3.2. *Text Preprocessing*

Text Preprocessing adalah pemrosesan teks yang berlangsung setelah tahap pemrosesan awal selesai. Pada *text preprocessing* sendiri terdapat beberapa tahap yang perlu dilakukan, seperti *text splitting, text cleansing, stopward removal,* dan *tokenizing.*

A. *Text Splitting*

Splitting adalah proses pembagian dataset atau teks menjadi beberapa kalimat-kalimat. *Text Splitting* merujuk pada proses pembagian teks yang lebih panjang menjadi lebih singkat dan terfokus [13].

B. *Text Cleansing*

Text Cleansing merupakan proses pembersihan atau pra-pemrosesan teks sebelum penerapan algoritma textrank yang mencakup penghapusan tanda baca, karakter khusus atau symbol yang tidak relevan dan merubah teks menjadi format yang konsisten (missalnya mengubah menjadi huruf kecil semua) [14].

C. *Stopward Removal*

Stopward removal adalah Kumpulan kata-kata umum yang sering muncul dalam teks namun tidak mempunyai makna atau nilai informasi yang tinggi dalam teks. Contoh *stopward removal* adalah “dan”, “atau”, “yang”, dan sebagainya dalam bahasa yang digunakan [13].

D. *Tokenizing*

Tokenizing adalah proses mengubah teks menjadi unit unit yang lebih kecil

dan lebih terfokus, *tokenizing* memungkinkan textrank untuk mengidentifikasi dan menganalisis struktur penting dari teks yang ada dengan memotong kata tunggal dengan (“ “) spasi sebagai delimeter [15].

3.3. TextRank

TextRank adalah algoritm pemrosesan bahasa alami yang digunakan untuk menganalisis teks dan memilih informasi penting dalam teks berdasarkan grafik teks. TextRank sendiri memiliki tahapan-tahapan umum dalam memilih informasi penting dari teks. Tahapan tersebut diantaranya *Text Similarity*, Pembobotan Teks, *Sorting*, dan Pemotongan Kalimat.

A. *Text Similarity*

Text Similarity adalah ukuran kesamaan atau sejauh mana dua atau lebih elemen teks memiliki kesamaan dalam isi, ataupun makna. *Text Similarity* dalam TextRank digunakan dalam proses rangkuman teks. Semakin tinggi kesamaan antara dua bagian teks, semakin besar pula kemungkinan informasi yang disajikan oleh kedua bagian tersebut dianggap sama pentingnya dalam ringkasan [16].

B. Pembobotan Teks

Pembobotan Teks merupakan proses pemberian nilai atau bobot pada kata dalam sebuah teks berdsarkan dari kriteria tertentu

C. *Sorting*

Sorting merupakan proses penyusunan kata dalam urutan tertentu berdasarkan suatu kriteria tertentu seperti frekuensi atau bobot tiap kata. *Sorting* sendiri digunakan dalam mengolah sebuah teks dengan cara mengidentifikasi, mengatur, dan mengevaluasi pentingnya sebuah kata.

D. Pemotongan Kalimat

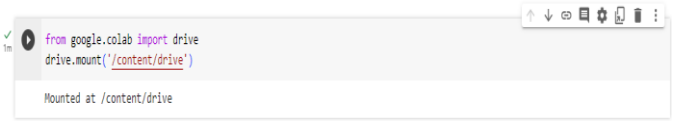
Pemotongan Kalimat merupakan sebuah proses menyusun teks panjang menjadi ringkasan yang lebih pendek dengan tetap menjaga esensi informasi yang diberikan.

4. HASIL DAN PEMBAHASAN

Hasil dan pembahasan untuk penerapan algoritma *textrank* dalam merangkum teks word dan pdf dilakukan menggunakan bantuan *Google Colab*. Untuk pengkoneksian antara *Google Colab* dengan *Google Drive* sebagai tempat pengambilan data, *penginstallan library*, pemanggilan data dari *Google Drive*, dan penggunaan algoritma *textrank* disajakan pada sub bab selanjutnya.

4.1. Pengkoneksian *Google Colab* dengan *Google Drive*

Pada tahap ini, melakukan pengkoneksian antara *Google Colab* dengan *Google Drive. Cara pengkoneksiannya sebagai berikut*



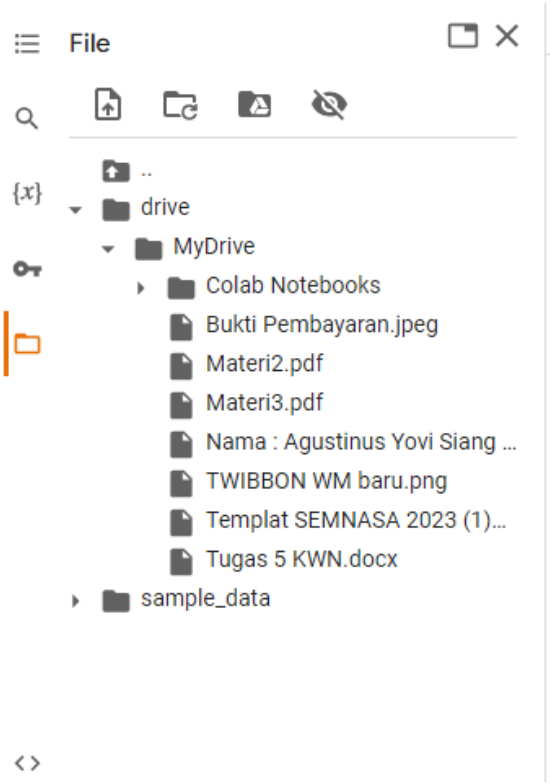
Gambar 2. Perintah pengkoneksian *Google Colab* dengan *Google Drive*

Gambar 2 berisi perintah untuk melakukan pengkoneksian antara *Google Colab* dengan *Google Drive.*

Gambar 3. Izin untuk penyambungan *Google Colab* dengan *Google Drive*

Gambar 4. Pilih akun *Google*

Gambar 3 dan Gambar 4 menunjukan tampilan untuk permintaan ijin mengkoneksikan *Google Colab* dengan *Google Drive*, serta pemilihan akun *Google* yang digunakan sebagai media penyimpanan data pada *Google Drivenya.*



Gambar 5.

Gambar 5 menunjukan tampilan bahwa *Google Colab* sudah terkoneksi dengan *Google Drive.*

4.2. *Penginstallan Library*

Selanjutnya adalah penginstallan paket *Phyton* dan Library yang digunakan untuk menjalankan *Textrank* pada *Google Colab.*

Gambar 6. Penginstallan paket *Phyton*

```
[6] from docx import Document
    from PyPDF2 import PdfReader
    from summa import summarizer
```

Gambar 7. Pengistallan *Library*

Pada gambar 7 merupakan *sourcode* yang digunakan untuk perintah mebaca dokumen dengan Pustaka *docx, PDF*, dan *sintax* untuk fungsi *summarizer* yang berguna untuk membuat ringkasan dengan menggunakan metode ekstraktif.

4.3. Pemanggilan Data Data dari *Google Drive*

```
[7] def read_docx(file_path):
        doc = Document(file_path)
        text = ''
        for para in doc.paragraphs:
            text += para.text
        return text
```

Gambar 8. Fungsi *Phyton* membaca dokumen dengan format *docx.*

```
def read_pdf(file_path):
    text = ''
    with open(file_path, 'rb') as file:
        pdf_reader = PdfReader(file)
        for page in pdf_reader.pages:
            text += page.extract_text()
    return text
```

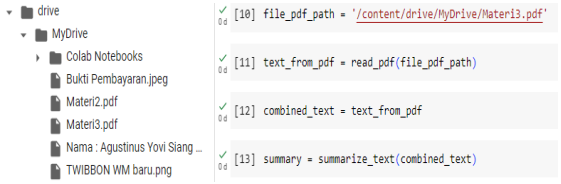
Gambar 9. Fungsi *Phyton* membaca file *PDF.*

```
def summarize_text(text):
    summary = summarizer.summarize(text, ratio=0.5)
    return summary
```

Gambar 10. Fungsi *Phyton* menggunakan metode *summarize*

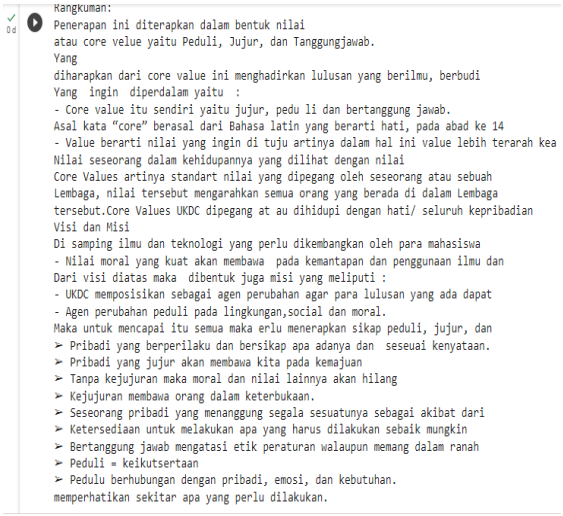
Pada gambar 10 merupakan fungsi untuk menerima sebuah teks sebagai argumen kemudian menggunakan *summarizer* untuk merangkum teks tersebut dengan menggunakan parameter *ratio = 0.5*

4.4. Penggunaan Algoritma *TextRank*



Gambar 11. Perintah Pemanggilan Data

Sebagai contoh pada Gambar 11 ditunjukan fungsi *Phyton* untuk pemanggilan data file *PDF* dari *Google Drive* yang telah terkoneksi, dengan nama file yang ada di *Google Drive* yaitu Materi3.pdf. Setelah melakukan pemanggilan data *PDF* kemudian dilakukan pembacaan dari file *PDF* yang dipanggil, yang selanjutnya setelah mengetahui isi dari file *PDF* yang dipanggil maka sistem akan melakukan pengambilan teks dan melakukan penyimpanan pada variabel *combined_text.* Lalu di tahap akhir melakukan peringkasan dengan perintah *summarize_text* dan mendapat hasil ringksan dari file *PDF* yang dipanggil seperti pada Gambar 12.



Gambar 12. Hasil ringkasan dari algoritma *textrank*

5. KESIMPULAN

Jurnal ini membahas penerapan algoritma Textrank dalam merangkum teks Word dan PDF serta penggunaan Google Colab dan Google Drive untuk menjalankan algoritma TextRank dalam membuat ringkasan teks. Teknologi saat ini memungkinkan pertumbuhan informasi yang cepat, namun juga memunculkan masalah kurangnya waktu untuk menyelidiki informasi lebih mendalam. Algoritma Textrank merupakan metode pemrosesan bahasa alami berbasis web yang menggunakan pendekatan unsupervised dan dapat digunakan untuk membuat peringkas teks otomatis. Metode penelitian yang digunakan meliputi text preprocessing, penggunaan algoritma Textrank, langkah-langkah pengkoneksian Google Colab dengan Google Drive, penginstallan library Python, pemanggilan data dari Google Drive, dan penggunaan algoritma TextRank. Daftar pustaka juga disertakan sebagai referensi.

UCAPAN TERIMA KASIH

Penulis mengucapkan terimakasih terhadap pihak-pihak yang telah mendukung untuk menyelesaikan penelitian ini.

DAFTAR PUSTAKA

[1] M. Danuri, "Development And Transformation Of Digital Technology," *Infokam*, Vol. Xv, No. Ii, Pp. 116–123, 2019.

[2] R. F. Daud, I. Komunikasi, U. M. Kotabumi, And L. Utara, "Dampak Perkembangan Teknologi Komunikasi Terhadap Bahasa Indonesia," *J. Interak. J. Ilmu Komun.*, Vol. 5, No. 2, Pp. 252–269, 2021, Doi: 10.30596/Interaksi.V5i2.7539.

[3] D. Andriani, I. Indriati, And M. T. Furqon, "Peringkasan Teks Otomatis Pada Artikel Berita Hiburan Berbahasa Indonesia Menggunakan Metode Bm25," *J. Pengemb. Teknol. Inf. Dan Ilmu Komput.*, Vol. 3, No. 3, Pp. 2603–2610, 2019.

[4] R. R. Putra, K. D. Evita, S. Pd, And M. Sc, "Automatic Summarization Text On Multi Document Using Textrank Faculty Of Engineering And Computer Science University Computer Indonesia."

[5] R. Mihalcea And P. Tarau, "Textrank: Bringing Order Into Texts."

[6] R. Restu Putra, "Peringkasan Teks Otomatis Pada Multi Dokumen Menggunakan Textrank." Universitas Komputer Indonesia, 2018.

[7] Y. Marsyah, "Perbandingan Kinerja Algoritme Textrank Dengan Algoritme Lexrank Pada Peringkasan Dokumen Bahasa Indonesia."

[8] Y. Liu And M. Lapata, "Text Summarization With Pretrained Encoders," *Arxiv Prepr. Arxiv1908.08345*, 2019.

[9] K. Ivanedra And M. Mustikasari, "Implementasi Metode Recurrent Neural Network Pada Text Summarization Dengan Teknik Abstraktif," *J. Teknol. Inf. Dan Ilmu Komput*, Vol. 6, No. 4, P. 377, 2019.

[10] A. N. Ammar And S. Suyanto, "Peringkasan Teks Ekstraktif Menggunakan Binary Firefly Algorithm," *Indones. J. Comput.*, Vol. 5, No. 2, Pp. 31–42, 2020.

[11] L. Pertiwi, "Penerapan Algoritma Text Mining, Steaming Dan Texrank Dalam Peringkasan Bahasa Inggris," 2022.

[12] N. Zhou, W. Shi, R. Liang, And N. Zhong, "Textrank Keyword Extraction Algorithm Using Word Vector Clustering Based On Rough Data-Deduction," *Comput. Intell. Neurosci.*, Vol. 2022, 2022.

[13] N. R. Siahaan, R. Y. Tiffany, S. R. E. Sinaga, E. V. N. B. Naibaho, And M. I. Fahmi, "Analisis Sentimen Ulasan Aplikasi Media Sosial Whatsapp Menggunakan Metode Naive Bayes Classifier," *J. Ilm. Betrik*, Vol. 14, No. 02 Agustus, Pp. 343–354, 2023.

[14] O. I. Gifari, M. Adha, F. Freddy, And F. F. S. Durrand, "Film Review Sentiment Analysis Using Tf-Idf And Support Vector Machine," *J. Inf. Technol.*, Vol. 2, No. 1, Pp. 36–40, 2022, Doi: 10.46229/Jifotech.V2i1.330.

[15] M. Munthe, "Perancangan Aplikasi Silogisme Artikel Bahasa Batak Dengan Menerapkan Metode Textrank," *J. Comput. Informatics Res.*, Vol. 1, No. 2, Pp. 34–37, 2022.

[16] S. Pratama And G. Alam, "Penerapan Algoritma Centroid-Based Summarization Untuk Sistem Peringkasan Dokumen Berbahasa Indonesia," *Julyxxxx*, Vol. X, No.X, Pp. 1–5.