

Penentuan Kalimat Utama Berbasis Pembobotan Kemiripan Makna

Ardi Sanjaya¹, Finsa Nurpandi², Marga Asta Jaya Mulya³

¹Teknik Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Nusantara PGRI Kediri

²Teknik Informatika, Universitas Suryakencana Cianjur

³Organisasi Riset Energi dan Manufaktur. Badan Riset dan Inovasi Nasional (BRIN), Tangerang

E-mail: ¹dersky@gmail.com, ²finsa@unsur.ac.id, ³marg002@brin.go.id

Abstrak – Identifikasi jenis paragraf berdasarkan letak kalimat utama merupakan bagian penting dalam analisis struktur teks bahasa Indonesia. Paragraf deduktif ditandai dengan kalimat utama di awal paragraf, sedangkan paragraf induktif memiliki kalimat utama di akhir paragraf. Proses identifikasi tersebut masih banyak dilakukan secara manual dan bersifat subjektif, sehingga diperlukan pendekatan otomatis berbasis *Natural Language Processing*. Penelitian ini mengusulkan metode deteksi kalimat utama untuk klasifikasi paragraf deduktif dan induktif dengan memanfaatkan pembobotan kemiripan makna menggunakan TF-IDF dan pengukuran kemiripan menggunakan *Cosine Similarity*. Setiap paragraf dipecah menjadi beberapa kalimat, kemudian dihitung nilai rata-rata kemiripan antar kalimat untuk menentukan kalimat utama berdasarkan nilai tertinggi. Pengujian dilakukan terhadap 220 paragraf yang terdiri dari 110 paragraf deduktif dan 110 paragraf induktif. Hasil evaluasi menunjukkan bahwa pada kelas paragraf deduktif diperoleh nilai *precision* sebesar 89,90%, *recall* 84,76%, dan *F1-score* 87,24%, sedangkan pada paragraf induktif diperoleh *precision* 88,04%, *recall* 79,41%, dan *F1-score* 83,49%. Kesalahan klasifikasi umumnya terjadi pada paragraf dengan struktur yang ambigu atau memiliki kalimat penutup atau pembuka yang bersifat rangkuman. Dapat disimpulkan bahwa pendekatan berbasis kemiripan makna antar kalimat efektif digunakan untuk mendeteksi kalimat utama dan mengklasifikasikan paragraf deduktif serta induktif dalam bahasa Indonesia.

Kata Kunci — NLP, TF-IDF, Kemiripan Makna, Kalimat Utama, *Cosine Similarity*

1. PENDAHULUAN

Struktur paragraf merupakan unsur penting dalam penulisan bahasa Indonesia karena mencerminkan keterpaduan gagasan dan kejelasan alur berpikir penulis. Berdasarkan letak kalimat utamanya, paragraf dapat diklasifikasikan menjadi paragraf deduktif, induktif, dan campuran [1][2][3]. Pada paragraf deduktif, gagasan utama disampaikan di awal paragraf, sedangkan pada paragraf induktif gagasan utama berada di akhir paragraf. Berdasar hasil studi literasi, identifikasi jenis paragraf dan penentuan kalimat utama selama ini masih banyak dilakukan secara manual, sehingga bersifat subjektif dan kurang efisien apabila diterapkan pada dokumen teks dalam jumlah besar.

Dalam bidang *Natural Language Processing* (NLP), pengukuran kemiripan teks merupakan salah satu pendekatan yang banyak digunakan untuk menganalisis hubungan antar kalimat. Metode *Cosine Similarity* dengan pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF) telah banyak diterapkan pada berbagai studi, seperti deteksi plagiarisme, penilaian jawaban esai, dan pengukuran kemiripan kalimat, karena mampu merepresentasikan teks dalam bentuk vektor numerik secara efektif [4][5][6]. Namun, sebagian besar penelitian tersebut masih menekankan pada kemiripan sintaksis, sehingga kalimat yang memiliki makna sama tetapi susunan kata berbeda sering kali menghasilkan nilai kemiripan yang rendah [6]. Penelitian sebelumnya mengusulkan pendekatan pengukuran kemiripan makna dengan memanfaatkan basis data sinonim kata, di mana kata-kata yang memiliki makna sama dikelompokkan dan direpresentasikan menggunakan ID unik. Pendekatan ini terbukti mampu meningkatkan nilai kemiripan antar kalimat yang secara semantik setara meskipun berbeda secara leksikal [7]. Konsep representasi kata berbasis ID sinonim ini membuka peluang untuk menganalisis hubungan makna antar kalimat dalam satu paragraf secara lebih mendalam.

Meskipun berbagai penelitian terkait pengukuran kemiripan kalimat telah banyak dilakukan, sebagian besar masih berfokus pada perbandingan antar dua kalimat secara terpisah, misalnya untuk deteksi plagiarisme atau penilaian kesamaan jawaban, dan belum diarahkan untuk menganalisis struktur internal sebuah paragraf. Penelitian-penelitian tersebut umumnya belum memanfaatkan distribusi nilai kemiripan antar seluruh kalimat dalam satu paragraf untuk menentukan peran kalimat utama. Selain itu, kajian yang secara khusus mengaitkan hasil pengukuran kemiripan makna dengan klasifikasi jenis paragraf deduktif dan induktif dalam bahasa Indonesia masih sangat terbatas. Oleh karena itu, diperlukan pendekatan yang tidak hanya mengukur kemiripan antar kalimat, tetapi juga mampu mengidentifikasi kalimat yang paling representatif terhadap isi paragraf secara keseluruhan sebagai dasar penentuan jenis paragraf.

Berdasarkan latar belakang tersebut, penelitian ini mengusulkan metode untuk mendeteksi jenis paragraf deduktif dan induktif dengan menentukan letak kalimat utama berdasarkan tingkat kemiripan makna antar kalimat dalam satu paragraf. Letak kalimat utama tersebut selanjutnya digunakan sebagai dasar klasifikasi paragraf. Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan analisis struktur paragraf berbasis NLP untuk bahasa Indonesia.

2. METODE PENELITIAN

2.1 Metode Penelitian

Penelitian ini menggunakan metode Waterfall sebagai pendekatan pengembangan sistem yang dilakukan secara bertahap dan terstruktur. Tahap awal diawali dengan studi literatur untuk mengkaji konsep struktur paragraf, kalimat utama, NLP, serta metode pengukuran kemiripan teks yang relevan, seperti TF-IDF dan Cosine Similarity. Selanjutnya dilakukan analisis kebutuhan untuk menentukan spesifikasi sistem, meliputi data masukan berupa paragraf, proses pemecahan kalimat dan pengolahan teks, serta keluaran sistem berupa posisi kalimat utama dan klasifikasi jenis paragraf.

Tahap berikutnya adalah perancangan sistem yang mencakup alur pemrosesan teks dan penentuan kalimat utama berdasarkan nilai rata-rata kemiripan antar kalimat. Sistem yang telah dirancang kemudian diuji menggunakan paragraf uji untuk mengetahui kinerjanya dalam mendeteksi kalimat utama dan mengklasifikasikan paragraf deduktif, induktif, atau bukan keduanya. Hasil pengujian selanjutnya dianalisis untuk mengevaluasi efektivitas metode yang diusulkan serta mengidentifikasi keterbatasan sebagai dasar penyusunan kesimpulan dan saran.

2.2 Landasan Teori

2.2.1 Paragraf Deduktif dan Induktif

Definisi paragraf menurut Dalman yang dikutip dari laporan penelitian yang berjudul Analisis Kesulitan Siswa Menemukan Kalimat Utama Dalam Teks Paragraf pada pelajaran Bahasa Indonesia Kelas V di SD Negeri 101832 Pancur Batu yaitu satuan pengembangan terkecil dari suatu karangan. Paragraf mengandung suatu pikiran pokok. Adapun kalimat-kalimat yang membentuk suatu paragraf umumnya dapat dibagi menjadi dua jenis, yaitu kalimat topik dan kalimat-kalimat jabatan. Definisi lain menurut R.Kunjana Rahardi dari sumber yang sama bahwa paragraf adalah satuan bahasa tulis yang terdiri dari beberapa kalimat. Kalimat-kalimat dalam paragraf itu harus disusun secara rutut dan sistematis. Sehingga dapat dijelaskan hubungan kalimat yang satu dengan kalimat lainnya dalam paragraf itu [8]. Suatu paragraf dikatakan deduktif apabila paragraf tersebut dimulai dengan mengemukakan persoalan pokok atau kalimat utama. Kemudian diikuti dengan kalimat-kalimat penjelas yang berfungsi menjelaskan kalimat utama. Paragraf ini biasanya dikembangkan dengan metode berfikir deduktif dari umum ke khusus. Sedangkan paragraf dikatakan sebagai induktif apabila paragraf ini dimulai dengan mengemukakan penjelasan-penjelasan atau perincian-perincian, kemudian ditutup dengan kalimat utama. Paragraf ini dikembangkan dengan metode berfikir induktif, dari hal-hal yang khusus ke umum [8].

2.2.2 Kalimat Utama

Menurut Wiyanto yang dikutip dari laporan penelitian yang berjudul Analisis Kesulitan Siswa Menemukan Kalimat Utama Dalam Teks Paragraf pada pelajaran Bahasa Indonesia Kelas V di SD Negeri 101832 Pancur Batu, kalimat utama atau kalimat topik adalah kalimat yang mengandung pokok pikiran sebuah paragraf. Kalimat utama dituangkan dalam satu kalimat diantara kalimat-kalimat lain yang terdapat pada sebuah paragraf yang menjadi pokok pikiran paragraf tersebut [8].

2.2.3 *Natural Language Processing*

Natural Language Processing (NLP) atau pemrosesan bahasa alami merupakan salah satu cabang ilmu AI yang berfokus pada pengolahan bahasa natural atau bahasa alami. Bahasa natural adalah bahasa yang secara umum digunakan oleh manusia dalam berkomunikasi satu sama lain [9]. NLP dalam arti luas untuk mencakup segala jenis manipulasi komputer terhadap bahasa alami. NLP melakukan proses pembuatan model komputasi dari bahasa, sehingga dapat terjadi suatu interaksi antara manusia dan komputer dengan perantaraan bahasa alami [10].

2.2.4 *Database Sinonim*

Database sinonim yang digunakan berasal dari penelitian sebelumnya [6]. Dengan mengacu pada *database* sinonim, akan dihasilkan suatu identitas unik (ID) yang nantinya digunakan sebagai dasar penentuan kemiripan makna tiap kata dan dasar pembobotan.



Gambar 1. Ilustrasi database sinonim

2.2.5 Preprocessing

Preprocessing adalah tahap awal dalam proses data mining yang bertujuan untuk meningkatkan kualitas data dengan cara membersihkan, mentransformasi, dan menyiapkan data agar siap digunakan pada tahap analisis atau pemodelan lebih lanjut [11]. Beberapa proses yang umum digunakan pada preprocessing diantaranya *case folding*, *tokenizing*, fungsi terbilang, *stopword removal* dan *stemming*.

2.2.6 Case Folding

Merupakan tahap perubahan huruf dari huruf kapital menjadi huruf kecil [12]. Hanya huruf a sampai z yang diterima. Karakter selain huruf dihilangkan dan dianggap delimitter [13]

2.2.7 Tokenizing

Tokenizing adalah proses memecah dokumen menjadi kumpulan kata [14]. *Tokenization* dapat dilakukan dengan menghilangkan tanda baca dan memisahkannya per spasi. Tahapan ini juga menghilangkan karakter-karakter tertentu seperti tanda baca dan mengubah semua token ke bentuk huruf kecil (*lower case*) [15].

2.2.8 Fungsi terbilang

Fungsi terbilang merupakan proses untuk mengubah angka menjadi suatu teks. Pada penelitian ini, fungsi terbilang digunakan untuk meningkatkan nilai kemiripan [7].

2.2.9 Stopword removal

Stopwords removal merupakan proses penghilangan kata tidak penting pada deskripsi melalui pengecekan kata-kata hasil parsing deskripsi apakah termasuk di dalam daftar kata tidak penting (stoplist) atau tidak. Jika termasuk didalam stoplist maka kata-kata tersebut akan dihapus dari deskripsi sehingga kata-kata yang tersisa di dalam deskripsi dianggap sebagai kata-kata penting atau keywords [15].

2.2.10 Stemming

Stemming merupakan proses untuk mencari kata dasar [15]. Proses *stemming* bahasa Indonesia adalah proses untuk mengeliminasi sufiks, prefiks, dan konfiks [16].

2.2.11 Cosine Similarity

Cosine Similarity mengukur kemiripan antara dua dokumen atau teks. Pada *Cosine Similarity* dokumen atau teks dianggap sebagai *vector*. Untuk pencocokan teks, nilai dari *vector* A dan B adalah *vektor term-frequency* dari dokumen. Nilai cosine similarity berada pada range 0-1 [4]. Persamaan *cosine similarity* disajikan pada persamaan 1 sebagai berikut :

$$(dj,q) = \frac{\sum_{i=l}^t (w_{ij}.w_{iq})}{\sqrt{\sum_{i=l}^t w_{ij}^2 \cdot \sum_{i=l}^t w_{iq}^2}} \dots\dots\dots (1)$$

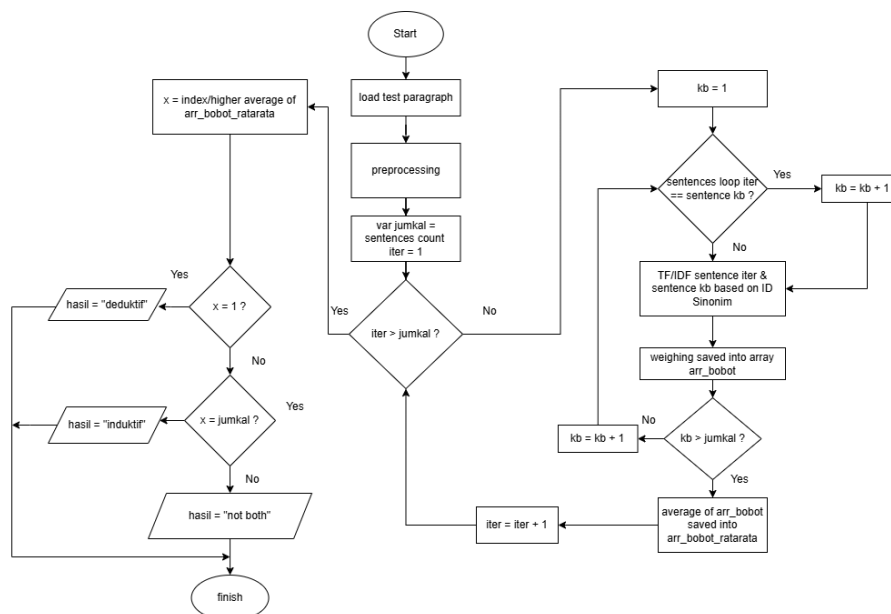
2.2.12 TF/IDF

Algoritma TF-IDF (*Term Frequency – Inverse Document Frequency*) adalah salah satu algoritma yang dapat digunakan untuk menganalisa hubungan antara sebuah frase/kalimat dengan sekumpulan dokumen atau frekuensi dari kemunculan sebuah term dalam dokumen yang bersangkutan.

2.3 Rancangan Aplikasi

Aplikasi dibangun menggunakan bahasa pemrograman Python. Aplikasi bekerja dengan memuat paragraf yang akan diuji. Kemudian dilakukan *preprocessing* dan memecah paragraf menjadi kalimat-kalimat. Misal pada

paragraf tersebut terdapat 4 kalimat. Pada masing-masing kalimat, setiap kata dikonversi menjadi ID. Pengkonversian menggunakan tokenizing dan ID Mapping. Pembobotan dengan TF/IDF pada kalimat 1 ke kalimat 2, kalimat 1 ke kalimat 3, kalimat 1 ke kalimat 4 (sebut saja iterasi 1) dan seterusnya sampai kalimat 4 ke kalimat 3. Hasil pembobotan kalimat 1 ke kalimat 2, kalimat 1 ke kalimat 3, kalimat 1 ke kalimat 4 kemudian dijumlah dan dirata-rata. Hasil pembobotan kalimat 2 ke kalimat 1, kalimat 2 ke kalimat 3, kalimat 2 ke kalimat 4 kemudian dijumlah dan dirata-rata dan seterusnya. Dari hasil rata-rata pembobotan diatas, dicari nilai tertinggi. Apabila Nilai tertinggi berada diiterasi 1 maka paragraf dapat disebut sebagai deduktif. Berikut disajikan alur program pada gambar 1.



Gambar 1. Alur program

2.4 Skenario Pengujian

Data yang digunakan berupa 110 paragraf deduktif dan 110 paragraf induktif. Total 220 paragraf tersebut dibuat oleh personal yang memiliki aktifitas dan spesialisasi dibidang Bahasa Indonesia dan telah diverifikasi oleh personal lainnya yang juga memiliki spesialisasi pada bidang linguistik Bahasa Indonesia.

Tabel 1. Contoh Data Paragraf

No	Paragraf	Label
1	Zaki seorang atlet yang multitalenta. Selain menjadi atlet olahraga voli, Zaki juga seorang fotografer yang sering pergi ke berbagai tempat yang indah untuk mengambil foto. Bakat yang dimiliki Zaki dalam olahraga voli, pernah mewakili juara Indonesia memenangkan Olimpiade pada tiga tahun yang lalu. Selain voli, Zaki juga masih sering bermain bola dengan teman sekampungnya. Bukan berarti dia tidak bisa dalam olahraga lain, justru dia mahir dalam beberapa bidang olahraga.	Deduktif
2	Kini banyak sekali pengembangan-pengembangan kewirausahaan sejak dini, mulai dari kalangan pelajar hingga kalangan mahasiswa. Dengan kemajuan teknologi informasi dan komunikasi sehingga mempermudah kita untuk melakukan bisnis dengan mudah. Banyak sekali para pemuda merencanakan bisnis yang keren, sehingga ketika ada modal langsung mengeksekusinya. Kini antusiasme pemuda terutama mahasiswa untuk menjadi pengusaha pun meningkat pesat	Induktif

3. HASIL DAN PEMBAHASAN

Berikut disajikan tabel 2 yang merupakan hasil pengujian terhadap 110 paragraf deduktif dan 110 paragraf induktif :

Tabel 2. Hasil pengujian

Ground Truth	Deduktif	Induktif	Bukan Keduanya	Error
Deduktif (110)	89	11	5	5
Induktif (110)	10	81	11	8

Pada pengujian terhadap 110 paragraf deduktif, sistem berhasil mengklasifikasikan 89 paragraf (80,91%) secara benar sebagai paragraf deduktif. Sebanyak 11 paragraf (10,00%) salah diklasifikasikan sebagai paragraf induktif dan 5 paragraf (4,55%) diklasifikasikan sebagai bukan deduktif maupun induktif. Selain itu, terdapat 5 paragraf (4,55%) yang tidak dapat diproses oleh sistem (*error*). Hasil ini menunjukkan bahwa sebagian besar paragraf deduktif dapat dikenali dengan baik oleh sistem, meskipun masih terdapat kesalahan klasifikasi yang umumnya disebabkan oleh kemiripan makna yang lebih dominan pada kalimat akhir paragraf.

Sementara itu, pada pengujian terhadap 110 paragraf induktif, sistem berhasil mengklasifikasikan 81 paragraf (73,64%) secara benar sebagai paragraf induktif. Sebanyak 10 paragraf (9,09%) terklasifikasi sebagai paragraf deduktif dan 11 paragraf (10,00%) terklasifikasi sebagai bukan deduktif maupun induktif. Selain itu, terdapat 8 paragraf (7,27%) yang tidak dapat diproses oleh sistem. Kesalahan klasifikasi pada paragraf induktif menunjukkan bahwa pada beberapa kasus, kalimat awal memiliki nilai kemiripan rata-rata yang relatif tinggi sehingga terdeteksi sebagai kalimat utama.

Tabel 2. Evaluasi performa

Kelas	Precision	Recall	F1-score
Deduktif	89,90%	84,76%	87,24%
Induktif	88,04%	79,41%	83,49%

Jika ditinjau secara keseluruhan dari 220 data paragraf, sistem berhasil mengklasifikasikan dengan benar sebanyak 170 paragraf, yang terdiri dari 89 paragraf deduktif dan 81 paragraf induktif. Dengan demikian, tingkat keberhasilan klasifikasi sistem terhadap dua jenis paragraf utama mencapai 77,27%. Adapun total kesalahan klasifikasi berjumlah 37 paragraf (16,82%), sedangkan 13 paragraf (5,91%) tidak dapat diproses akibat kendala teknis. Hasil ini menunjukkan bahwa pendekatan berbasis kemiripan makna antar kalimat cukup efektif untuk mendeteksi struktur paragraf deduktif dan induktif, meskipun masih memiliki keterbatasan pada paragraf dengan struktur semantik yang ambigu.

4. SIMPULAN

Berdasarkan hasil pengujian dan evaluasi performa menggunakan precision, recall, dan F1-score, metode yang diusulkan menunjukkan kinerja yang cukup baik dalam mengklasifikasikan paragraf deduktif dan induktif. Pada kelas paragraf deduktif, sistem memperoleh nilai precision sebesar 89,90%, recall sebesar 84,76%, dan F1-score sebesar 87,24%, yang menunjukkan kemampuan sistem dalam mengenali pola deduksi dengan tingkat ketepatan dan konsistensi yang tinggi. Sementara itu, pada kelas paragraf induktif, sistem menghasilkan precision sebesar 88,04%, recall sebesar 79,41%, dan F1-score sebesar 83,49%. Nilai recall yang relatif lebih rendah pada paragraf induktif menunjukkan bahwa sebagian paragraf induktif masih terdeteksi sebagai deduktif atau bukan keduanya, terutama pada paragraf yang memiliki kalimat awal dengan tingkat kemiripan makna yang tinggi. Secara keseluruhan, hasil ini menunjukkan bahwa pendekatan berbasis kemiripan makna antar kalimat efektif digunakan untuk analisis struktur paragraf, meskipun masih memiliki keterbatasan pada paragraf dengan struktur yang bervariasi.

5. SARAN

Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa saran yang dapat dijadikan acuan untuk pengembangan penelitian selanjutnya. Pertama, penelitian ini masih menunjukkan adanya kesalahan klasifikasi, terutama pada paragraf induktif yang memiliki struktur semantik ambigu atau kalimat pembuka yang bersifat rangkuman. Oleh karena itu, penelitian selanjutnya disarankan untuk menambahkan fitur pendukung, seperti bobot posisi kalimat, atau panjang kalimat guna meningkatkan kemampuan sistem dalam membedakan kalimat utama dan kalimat penjelas secara lebih akurat. Kedua, masih ditemukannya data yang tidak dapat diproses oleh sistem menunjukkan perlunya peningkatan pada tahap pra-pemrosesan teks. Untuk penelitian selanjutnya diharapkan dapat mengembangkan mekanisme penanganan data tidak valid, seperti paragraf dengan jumlah kalimat sangat sedikit atau struktur kalimat yang tidak lengkap, agar sistem mampu menangani variasi data teks yang lebih luas.

DAFTAR PUSTAKA

- [1] Z. A.-F. Aguslim, Cut Marlini, “Analisis Kemampuan Siswa Menentukan Kalimat Utama dalam Pembelajaran Bahasa Indonesia Kelas 4 SD Negeri 49 Kota Banda Aceh,” *J. Ilm. Mhs. Univ. Bina Bangsa Getsempena*, vol. 2, no. 1, 2021, [Online]. Available: <https://jim.bbg.ac.id/pendidikan/article/download/323/160/431>
- [2] M. Khairat, “IMPLEMENTASI K-MEANS CLUSTERING DAN VECTOR SPACE MODEL UNTUK MENCARI KALIMAT UTAMA TEKS ARTIKEL BERBAHASA INDONESIA,” Universitas Islam Negeri Sultan Syarif Kasim Riau, 2019.
- [3] D. A. R. Ariantini, A. S. M. Lumenta, and A. Jacobus, “Pengukuran Kemiripan Dokumen Teks Bahasa Indonesia Menggunakan Metode Cosine Similarity,” *J. Tek. Inform.*, vol. 9, no. 1, pp. 1–8, 2016, doi: 10.35793/jti.9.1.2016.13752.
- [4] R. P. Pratama, M. Faisal, and A. Hanani, “Deteksi Plagiarisme pada Dokumen Jurnal Menggunakan Metode Cosine Similarity,” *SMARTICS J.*, vol. 5, no. 1, pp. 22–26, 2019, doi: 10.21067/smartics.v5i1.2848.
- [5] E. L. Amalia, A. J. Jumadi, I. A. Mashudi, and D. W. Wibowo, “Analisis Metode Cosine Similarity Pada Aplikasi Ujian Online Otomatis (Studi Kasus JTI POLINEMA),” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 2, p. 343, 2021, doi: 10.25126/jtiik.2021824356.
- [6] A. Sanjaya, A. B. Setiawan, U. Mahdiyah, I. N. Farida, and A. R. Prasetyo, “Pengukuran Kemiripan Makna Menggunakan Cosine Similarity Measurement of Meaning Similarity Using Cosine Similarity and,” vol. 10, no. 4, pp. 747–752, 2023.
- [7] A. Sanjaya and S. D. Sasongko, “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,” vol. 7, no. 2, pp. 95–104, 2022.
- [8] F. Sibarani, “ANALISIS KESULITAN SISWA MENEMUKAN KALIMAT UTAMA DALAM TEKS PARAGRAF PADA MATA PELAJARAN BAHASA INDONESIA KELAS V DI SD NEGERI 101832 PANCUR BATU TAHUN AJARAN 2019/2020,” Universitas Quality, 2020. [Online]. Available: <http://portaluniversitasquality.ac.id:55555/949/7/BAB II.pdf>
- [9] D. Suhartono, “Natural Language Processing.” Accessed: Feb. 07, 2022. [Online]. Available: <https://socs.binus.ac.id/2013/06/22/natural-language-processing/>
- [10] E. L. Steven Bird, Ewan Klien, *Natural Language Processing with Python*, First Edit. Sebastopol, CA 95472.: O’reilly Media Inc, 2009. Accessed: Feb. 07, 2022. [Online]. Available: <http://www.datascienceassn.org/sites/default/files/Natural Language Processing with Python.pdf>
- [11] J. Han, J.W., Kamber, M. and Pei, *Data Mining: Concepts and Techniques (3rd ed.)*. Waltham: Morgan Kaufmann Publishers, 2012.
- [12] I. Mawanta, T. S. Gunawan, and W. Wanayumini, “Uji Kemiripan Kalimat Judul Tugas Akhir dengan Metode Cosine Similarity dan Pembobotan TF-IDF,” *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 726, 2021, doi: 10.30865/mib.v5i2.2935.
- [13] M. A. Salim and Y. Anistyasari, “Pengembangan Aplikasi Penilaian Ujian Essay Berbasis Online Menggunakan Algoritma Nazief Dan Adriani Dengan Metode Cosine Mohammad Agus Salim Yeni Anistyasari Abstrak,” *IT-Edu J. Inf. Technol. Educ.*, vol. 02, no. 1, pp. 126–135, 2017.
- [14] M. M. umilasari Syabani reni, “Penerapan Metode Cosine Similarity dan Pembobotan TF/IDF pada Sistem Klasifikasi Sinopsis Buku di Perpustakaan Kejaksaan Negeri Jember,” *JUSTINDO (Jurnal Sist. dan Teknol. Inf. Indones.)*, no. Vol 3, No 1 (2018): JUSTINDO, pp. 31–42, 2018, [Online]. Available: <http://jurnal.unmuhjember.ac.id/index.php/JUSTINDO/article/view/2345>
- [15] K. S. Nugroho, “Dasar Text Preprocessing dengan Python.” Accessed: Feb. 07, 2022. [Online]. Available: <https://ksnugroho.medium.com/dasar-text-preprocessing-dengan-python-a4fa52608ffe>
- [16] A. Librian, “High quality stemmer library for Indonesian Language (Bahasa).” Accessed: Feb. 08, 2022. [Online]. Available: <https://github.com/sastrawi/>