**Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)** DOI: 10.25126/jtiik.2023106864
Vol. 10, No. 4, Agustus 2023, hlm. 747-752 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

# PENGUKURAN KEMIRIPAN MAKNA MENGGUNAKAN COSINE SIMILARITY DAN BASIS DATA SINONIM KATA

**Ardi Sanjaya[*1], Ahmad Bagus Setiawan[2], Umi Mahdiyah[3], Intan Nur Farida[4], Aprisa Risky Prasetyo[5]**

[1,2,3,4,5] Universitas Nusantara PGRI Kediri, Kediri
Email: [1]dersky@gmail.com, [2]bagus.este@gmail.com, [3]umimahdiyah@gmail.com, [4]in.infarida@gmail.com, [5]aprisariskyprasetyo2442@gmail.com

(Naskah masuk: 29 Desember 2023, diterima untuk diterbitkan: 25 Juli 2023)

**Abstrak**

Penelitian ini bertujuan untuk memberikan alternatif dalam menguji kemiripan makna antar 2 kalimat. Pembentukan database sinonim kata dilakukan dengan mengelompokkan kata berdasar sinonim atau yang memiliki kesamaan arti. Masing-masing kelompok kata diberikan ID unik. Selanjutnya setiap kelompok kata dipecah untuk diuraikan menjadi kata tunggal, disimpan pada tabel kata dengan melabeli ID kata dan ID sinonim. ID sinonim didasarkan pada ID unik pada tabel sinonim. Dalam pengujian kemiripan makna, masing-masing kalimat akan di urai menjadi kata dan tiap-tiap kata akan dicocokkan berdasarkan tabel kata dengan acuan ID sinonim. ID Sinonim yang didapat kemudian dilakukan pengukuran jarak vektor dan kemiripan menggunakan rumus cosine similarity. Berdasarkan pengujian dan analisa yang telah dilakukan, dari 25 pengujian didapati 24 nilai kemiripan mengalami peningkatan prosentase. Hal tersebut dikarenakan penggunaan ID yang didasarkan pada kelompok kata dan irisan saat proses pembobotan mampu meningkatkan nilai kemiripan. Rata-rata nilai kemiripan pada penggunaan ID sebagai vektor hitung adalah 94,48% dan rata-rata nilai kemiripan pada metode atau alur pembanding adalah sebesar 69,96%.

**Kata kunci**: *kemiripan makna, cosine similarity, sinonim bahasa Indonesia*

# *MEASUREMENT OF MEANING SIMILARITY USING COSINE SIMILARITY AND WORD SYNONYMS DATABASE*

***Abstract***

*This study aims to provide an alternative in testing the similarity of meaning between 2 sentences. The formation of a word synonym database is done by grouping words based on synonyms or those that have the same meaning. Each group of words is assigned a unique ID. Furthermore, each group of words is broken down to be broken down into single words, stored in the word table labeled word ID and synonym ID. Synonym ID is based on the unique ID in the synonym table. In testing the similarity of meaning, each sentence will be broken down into words and each word will be matched based on the word table with synonym ID references. The synonym ID obtained is then measured by measuring the vector distance and similarity using the cosine similarity formula. Based on the tests and analyzes that have been carried out, out of 25 tests it was found that 24 similarity values experienced an increase in the percentage. This is because the use of ID based on word groups and slices during the weighting process can increase the similarity value. The average similarity value in the use of ID as a calculating vector is 94.48% and the average similarity value in the comparison method or plot is 69.96%.*

**Keywords**: *similarity of meaning, cosine similarity, Indonesian*

## 1. PENDAHULUAN

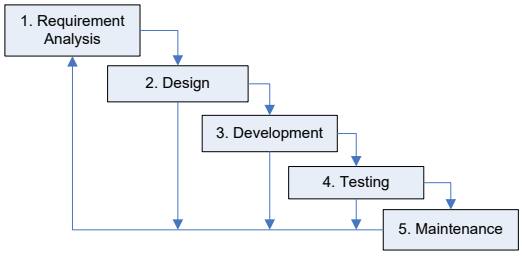
Pengukuran kemiripan antar 2 kalimat masih menjadi topik penelitian terutama sejak masa pandemi Covid 19. Salah satunya bisa diterapkan untuk mengukur kemiripan antara jawaban peserta didik dan kunci jawaban pada *Learning Manajemen System* (LMS). Berdasarkan pengamatan penulis, kebanyakan pengukuran kemiripan kalimat tersebut hanya mengukur sebatas kemiripan secara sintaksis (penulisan) yang diubah ke bentuk dasar. Sedangkan yang meneliti untuk mengukur kemiripan makna masih sedikit. (Abriani & Yaqin, 2019) dalam penelitiannya yang berjudul Analisis Implementasi Metode Semantic Similarity untuk Pengukuran Kemiripan Makna Antar Kalimat menggunakan sentence similarity untuk menghitung kemiripan antar kata dan metode Analytical Hierrarchy Process untuk menentukan bobot kriteria noun dan verb. Penelitian yang terdahulu berjudul Pengukuran

Kemiripan Makna Kalimat Dalam Bahasa Indonesia Menggunakan pendekatan path karena paling sesuai untuk menghitung jumlah node atau relasi yang terhubung antar node lain dalam sinonimkata.com (Caterina et al., 2021). Hasil yang didapat yaitu kemiripan kalimat dalam bahasa Indonesia yang memiliki tingkat kemiripan yang tinggi bernilai 0,875 pada eksperimen kriteria kalimat dengan susunan kata kerja – kata benda. (Amalia et al., 2021) dalam penelitiannya menggunakan Cosine similarity dan TF untuk menilai kemiripan pada studi kasusnya pembuatan aplikasi ujian online easi otomatis. Hasil pengujian precision, recall dan f-measure diperoleh rata-rata 80%.

Dari beberapa referensi, terdapat pengujian dimana pada kalimat dengan kata berbeda namun makna sama memiliki nilai kemiripan rendah. Gambaran dari penelitian ini adalah peneliti mengelompokkan kata berdasarkan makna yang sama dengan mengacu pada database sinonim dari Kateglo. Selanjutnya pengukuran kemiripan dilakukan tiap kata dengan mengaju kepada ID unik masing-masing kelompok sinonim tadi. Apabila dua kata yang dibandingkan memiliki sintak berbeda namun memiliki ID kelompok sinonim sama maka kedua kata tersebut akan dinilai sama. Penelitian ini merupakan lanjutan dan pengembangan dari penelitian-penelitian sebelumnya terkait pengukuran kemiripan kalimat dan makna serta bertujuan untuk memberikan alternatif untuk mengukur kemiripan makna dalam bahasa Indonesia.

## 2. METODE PENELITIAN

Dalam membangun sistem pada penelitian ini, menggunakan metode Software Life Cycle Developmet (SLCD) tahapan-tahapan seperti tersaji pada gambar berikut :




Gambar 1. Siklus pembuatan sistem

Tahap awal yaitu menganalisa kebutuhan, dilanjutkan dengan desain sistem, pembuatan program (developmet), pengujian dan maintenance.

### 2.1 Landasan Teori

#### 2.1.1 Natural Langauge Processing

Natural Language Processing (NLP) 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 (Suhartono, 2013). 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 (Steven Bird, Ewan Klien, 2009).

#### 2.1.2 Case Folding

Merupakan tahap perubahan huruf dari huruf kapital menjadi huruf kecil (Mawanta et al., 2021). Hanya huruf a sampai z yang diterima. Karakter selain huruf dihilangkan dan dianggap delimiter (Salim & Anistyasari, 2017).

#### 2.1.3 Tokenizing

Tokenizing adalah proses memecah dokumen menjadi kumpulan kata (Syabani reni, 2018). 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*) (Nugroho, 2019).

#### 2.1.4 Fungsi Terbilang

Fungsi terbilang merupakan proses untuk mengubah angka menjadi suatu teks. Pada penelitian ini, fungsi terbilang digunakan untuk meningkatkan nilai kemiripan (Sanjaya & Sasongko, 2022).

#### 2.1.5 Filtering/Stopwords 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 (Nugroho, 2019).

#### 2.1.6 Stemming

Stemming merupakan proses untuk mencari kata dasar (Nugroho, 2019). Proses *stemming* bahasa Indonesia adalah proses untuk mengeliminasi sufiks, prefiks, dan konfiks (Librian, 2017).

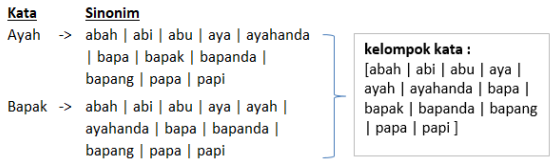
#### 2.1.7 Cosine Similarity

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 (Pratama et al., 2019). Persamaan cosine similarity disajikan pada persamaan 1 sebagai berikut :

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

## 2.2 Perancangan Basis Data Sinonim

Data sinonim kata yang digunakan pada penelitian ini menggunakan basis data sinonim dari aplikasi Kateglo yang dimodifikasi dan ditambahkan. Yang dibutuhkan pada penelitian ini adalah adanya pemberian ID pada kumpulan kata yang memiliki makna sama seperti disajikan pada ilustrasi gambar 2 berikut :




Gambar 2. Ilustrasi Pembentukan Kelompok Kata

Pada gambar diatas, kata ayah memiliki memiliki sinonim/persamaan makna dengan kata abah, abi, abu, aya, ayahanda, bapa, bapak dan seterusnya. Demikian juga kata bapak memiliki sinonim dengan kata abah, abi, abu, ayah dan seterusnya. Dalam penelitian ini, kata ayah dan bapak dapat dikelompokkan atau dikumpulkan menjadi satu dan di beri sebuah identitas atau ID. Sehingga dapat dirumuskan rancangan basis data dan tabel sinonim untuk kebutuhan tersebut.

Tabel 1. Rancangan tabel tbsinonim

| No | Kolom | Type | Keterangan |
|---|---|---|---|
| 1 | id_sinonim | Integer(11) | Primary key, auto increment |
| 2 | kumpulankata | text | |

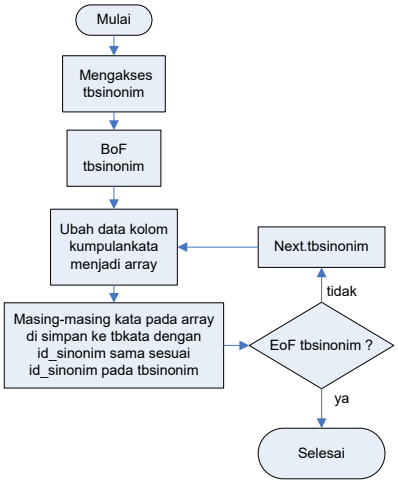
Agar lebih mudah dalam implementasi query, maka nantinya setelah data kumpulan kata selesai dibuat dan melabeli dengan ID sinonim, selanjutnya memisah masing-masing kata pada kolom kumpulankata dan melabeli dengan ID sinonim. Pemisahan tersebut ditunjukkan pada gambar 3 dan data kata ditampung pada tabel tbkata dengan struktur seperti tersaji pada tabel 2.

Tabel 2. Rancangan tabel tbkata

| No | Kolom | Type | Keterangan |
|---|---|---|---|
| 1 | Id_kata | Integer(11) | PK, auto increment |
| 2 | id_sinonim | Integer(11) | FK |
| 3 | kata | text | |

## 2.3 Perancangan Sistem/Aplikasi

Tahap pertama dari jalannya sistem adalah membaca input 2 kalimat yang akan diukur nilai kemiripannya. Misal kalimat pertama adalah "Ibu memasak nasi" Dan kalimat kedua adalah "Mama menanak nasi".

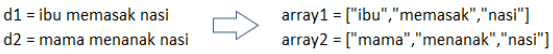



Gambar 3. Alur pengisian tabel tbkata

Kemudian masing-masing kalimat diubah ke *lower case* dan membuang tanda baca yang tidak perlu seperti tanda titik, koma dan sejenisnya. Proses selanjutnya yaitu memastikan kedua kalimat harus dalam bentuk kalimat aktif. Proses *lower case* diilustrasikan seperti pada gambar 4 berikut :

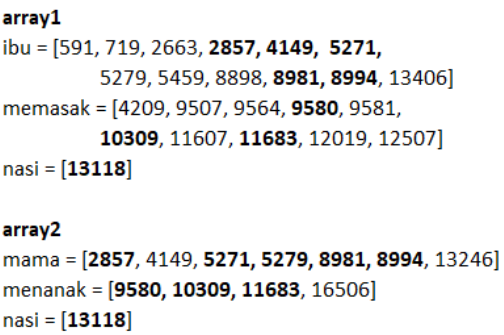

Gambar 4. Ilustrasi proses *lower case*

Kemudian masing-masing kalimat diurai menjadi kata dan di simpan pada array. Ilustrasinya disajikan pada gambar 5 berikut :

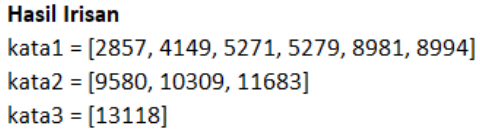



Gambar 5. Ilustrasi proses pengubahan menjadi array

Lalu pada masing-masing data array dilakukan konversi menjadi ID sinonim dengan cara mencocokan terhadap database sinonim. Gambaran proses tersebut disajikan pada gambar 6 berikut :




Gambar 7. Ilustrasi konversi dari array menjadi ID sinonim




Gambar 6. Ilustrasi hasil proses irisan ID

Setelah itu apabila jumlah anggota 2 array tersebut sama, maka dilakukan *intersect* (mencari data ID sinonim yang sama) pada kedua array tersebut. Apabila jumlah anggota tidak sama, maka dilakukan 2 kali proses intersect yaitu data array 1 terhadap data array 2 dan array 2 terhadap array 1. Kemudian dilakukan update pada masing-masing data array karena ada kemungkinan data ID sinonim yang sama ada lebih dari 1. Data ID sinonim yang sama hanya cukup 1 yang mewakili pada setiap kata.

**array1 =** [2857,9580,13118]
kata1 = 2857
kata2 = 9580
kata3 = 13118

**array2 =** [2857,9580,13118]
kata1 = 2857
kata2 = 9580
kata3 = 13118

Gambar 7. Ilustrasi proses penyatuan irisan pada array

Selanjutnya menghitung nilai kemiripan menggunakan consine similarity. Berdasar ilustrasi proses konversi menjadi ID sinonim diatas, maka dapat dihitung sebagai berikut:

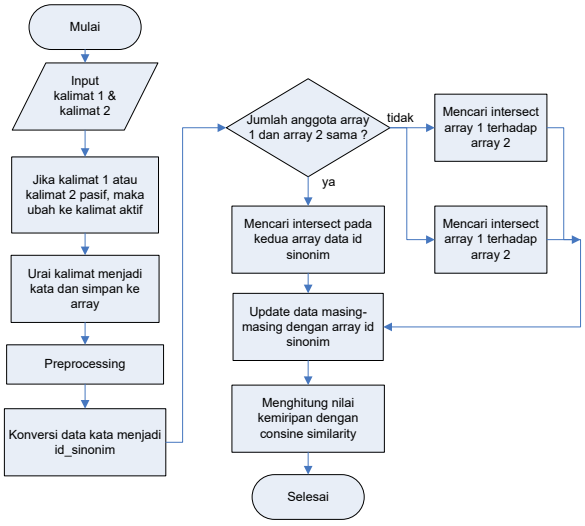
Tabel 3. Bobot konversi

| ID Sinonim | Bobot Array1 | Bobot Array2 |
|---|---|---|
| 2857 | 1 | 1 |
| 9580 | 1 | 1 |
| 13118 | 1 | 1 |

$$\cos(\emptyset) = \frac{(1x1)+(1x1)+(1x1)}{\sqrt{(1^2\ x\ 1^2\ )+(1^2\ x\ 1^2\ )+(1^2\ x\ 1^2\ )}} = 1$$

Nilai kemiripan = 1 x 100% = 100%.

Alur sistem atau aplikasi yang telah diilustrasikan diatas dan yang digunakan pada penelitian ini disajikan pada gambar 8 berikut:




Gambar 8. Alur sistem atau aplikasi

## 3. HASIL DAN PEMBAHASAN

Berikut disajikan screenshot pada gambar 9 struktur tabel tbsinonim dan gambar 10 struktur tabel tbkata yang telah dibuat

```
MariaDB [sinonim]> desc tbsinonim;
+-------------+---------+------+-----+---------+----------------+
| Field       | Type    | Null | Key | Default | Extra          |
+-------------+---------+------+-----+---------+----------------+
| id_sinonim  | int(11) | NO   | PRI | NULL    | auto_increment |
| kumpulankata| text    | NO   |     | NULL    |                |
+-------------+---------+------+-----+---------+----------------+
2 rows in set (0.042 sec)
```

Gambar 9. Screenshoot struktur tabel tbsinonim

```
MariaDB [sinonim]> desc tbkata;
+-----------+---------+------+-----+---------+----------------+
| Field     | Type    | Null | Key | Default | Extra          |
+-----------+---------+------+-----+---------+----------------+
| id_kata   | int(11) | NO   | PRI | NULL    | auto_increment |
| id_sinonim| int(11) | NO   |     | NULL    |                |
| kata      | text    | NO   |     | NULL    |                |
+-----------+---------+------+-----+---------+----------------+
3 rows in set (0.028 sec)
```

Gambar 10. Screenshoot struktur tabel tbkata

Sebagai pembanding dalam pengujian, peneliti menggunakan alur yang digunakan pada penelitian sebelumnya (Sanjaya & Sasongko, 2022).

Tabel 4. Data hasil pengukuran

| No | Pembanding (%) | Pengukuran (%) | Keterangan |
|---|---|---|---|
| 1 | 28.86 | 86.60 | Ada Peningkatan |
| 2 | 50 | 75.00 | Ada Peningkatan |
| 3 | 61.72 | 92.58 | Ada Peningkatan |
| 4 | 66.67 | 100.00 | Ada Peningkatan |
| 5 | 50.7 | 84.52 | Ada Peningkatan |
| 6 | 83.33 | 100.00 | Ada Peningkatan |
| 7 | 92.31 | 100.00 | Ada Peningkatan |
| 8 | 66.67 | 100.00 | Ada Peningkatan |
| 9 | 80 | 100.00 | Ada Peningkatan |
| 10 | 80 | 100.00 | Ada Peningkatan |
| 11 | 85.71 | 100.00 | Ada Peningkatan |
| 12 | 80 | 100.00 | Ada Peningkatan |
| 13 | 73.03 | 91.29 | Ada Peningkatan |
| 14 | 80 | 80.00 | Sama |
| 15 | 85.71 | 100.00 | Ada Peningkatan |
| 16 | 50 | 100.00 | Ada Peningkatan |
| 17 | 66.67 | 100.00 | Ada Peningkatan |
| 18 | 33.33 | 100.00 | Ada Peningkatan |
| 19 | 66.82 | 93.54 | Ada Peningkatan |
| 20 | 70 | 90.00 | Ada Peningkatan |
| 21 | 88.34 | 93.54 | Ada Peningkatan |
| 22 | 71.42 | 100.00 | Ada Peningkatan |
| 23 | 80 | 100.00 | Ada Peningkatan |
| 24 | 77.78 | 100.00 | Ada Peningkatan |
| 25 | 80 | 100.00 | Ada Peningkatan |

Berdasarkan hasil pengukuran kemiripan yang dilakukan seperti tersaji pada tabel 4 diatas, nilai minimum yang didapat adalah sebesar 75.00%, nilai maksimal 100.00% dan rata-rata sebesar 95,48%. Nilai minimum terjadi pada pengujian data kedua. Kalimat yang digunakan adalah sebagai berikut :

D1 = Andi telah menuntaskan pekerjaan
D2 = Andi berhasil menyelesaikan pekerjaan

Apabila dihitung menggunakan alur pembanding, didapatkan hasil:

Tabel 5. Pembobotan data ke dua pengujian pembanding

| Kata (Term) | TF | |
|---|---|---|
| | D1 | D1 |
| andi | 1 | 1 |
| telah | 1 | 0 |
| tuntas | 1 | 0 |
| pekerjaan | 1 | 1 |
| hasil | 0 | 1 |
| selesai | 0 | 1 |

D0 = [1,1,1,1,0,0]
D1 = [1,0,0,1,1,1]

Berdasar rumus cosine similarity, maka vektor D0 dan D1 memiliki jarak 0.5 dan nilai kemiripan sebesar 50,0%. Sedangkan pada alur yang digunakan peneliti, berikut analisa pada data ke dua:

Tabel 6. Pembobotan data ke dua

| Kata (Term) | ID | TF | |
|---|---|---|---|
| | | D0 | D1 |
| Andi | 888 | 1 | 1 |
| Telah | 6133 | 1 | 0 |
| Menuntaskan/menyelesaikan | 9458 | 1 | 1 |
| Pekerjaan | 227 | 1 | 1 |
| berhasil | 2154 | 0 | 1 |

Pada tabel 6 diatas, kata/term andi berdasar database tabel tbkata miliki ID 888, menuntaskan atau menyelesaikan memiliki irisan ID yang sama yaitu 9458, pekerjaan memiliki ID 227, dan berhasil memiliki ID 2154. Kemudian ditentukan vektor sebagai berikut:

D0 = [1,1,1,1,0]
D1 = [1,0,1,1,1]

Berdasar rumus cosine similarity, maka vektor D0 dan D1 memiliki jarak 0.75 dan nilai kemiripan sebesar 75,0%

## 4. KESIMPULAN DAN SARAN

Berdasarkan pengujian dan analisa yang telah dilakukan, didapati 24 nilai kemiripan dari 25 pengujian mengalami peningkatan. Hal tersebut dikarenakan penggunaan ID yang didasarkan pada kelompok kata dan irisan saat proses pembobotan mampu meningkatkan nilai kemiripan. Rata-rata nilai kemiripan pada penggunaan ID sebagai vektor hitung adalah 94,48% dan rata-rata nilai kemiripan pada metode atau alur pembanding adalah sebesar 69,96%.

## DAFTAR PUSTAKA

ABRIANI, G. U., & YAQIN, M. A., 2019. Implementasi Metode Semantic Similarity untuk Pengukuran Kemiripan Makna antar Kalimat. *ILKOMNIKA: Journal of Computer Science and Applied Informatics*, *1*(2), 47–57. https://doi.org/10.28926/ilkomnika.v1i2.15

AMALIA, E. L., JUMADI, A. J., MASHUDI, I. A., & WIBOWO, D. W., 2021. Analisis Metode Cosine Similarity Pada Aplikasi Ujian Online Otomatis (Studi Kasus JTI POLINEMA). *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *8*(2), 343. https://doi.org/10.25126/jtiik.2021824356

CATERINA, Y., YAQIN, M. A., & ZAMAN, S., 2021. Pengukuran Kemiripan Makna Kalimat dalam Bahasa Indonesia Menggunakan Metode Path. *Fountain of Informatics Journal*, *6*(2), 45. https://doi.org/10.21111/fij.v6i2.4844

LIBRIAN, A., 2017. *High quality stemmer library for Indonesian Language (Bahasa)*. https://github.com/sastrawi/

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

NUGROHO, K. S., 2019. *Dasar Text Preprocessing dengan Python*. https://ksnugroho.medium.com/dasar-text-preprocessing-dengan-python-a4fa52608ffe

PRATAMA, R. P., FAISAL, M., & HANANI, A., 2019. Deteksi Plagiarisme pada Dokumen Jurnal Menggunakan Metode Cosine Similarity. *SMARTICS Journal*, *5*(1), 22–26. https://doi.org/10.21067/smartics.v5i1.2848

SALIM, M. A., & ANISTYASARI, Y., 2017. Pengembangan Aplikasi Penilaian Ujian Essay Berbasis Online Menggunakan Algoritma Nazief Dan Adriani Dengan Metode Cosine Mohammad Agus Salim Yeni Anistyasari Abstrak. *IT-Edu : Jurnal Information Technology and Education*, *02*(1), 126–135.

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*. *7*(2), 95–104.

STEVEN BIRD, EWAN KLIEN, E. L., 2009. *Natural Language Processing with Python* (J. Steele (ed.); First Edit). O'reilly Media Inc. http://www.datascienceassn.org/sites/default/files/Natural Language Processing with Python.pdf

SUHARTONO, D., 2013. *Natural Language Processing*. https://socs.binus.ac.id/2013/06/22/natural-language-processing/

SYABANI RENI, M. M. UMILASARI., 2018. Penerapan Metode Cosine Similarity dan Pembobotan TF/IDF pada Sistem Klasifikasi Sinopsis Buku di Perpustakaan Kejaksaan Negeri Jember. *JUSTINDO (Jurnal Sistem Dan Teknologi Informasi Indonesia)*, *Vol 3*, *No 1 (2018):* *JUSTINDO*, 31–42. http://jurnal.unmuhjember.ac.id/index.php/JUSTINDO/article/view/2345