TESIS

KONSISTENSI PENULISAN JURNAL ILMIAH BERBASIS NATURAL LANGUAGE PROCESSING

CONSISTENCY OF SCIENTIFIC JOURNAL WRITING BASED ON NATURAL LANGUAGE PROCESSING

Disusun dan diajukan oleh

SITTI MAWADDAH UMAR

D082211023

PROGRAM STUDI MAGISTER TEKNIK INFORMATIKA

DEPARTEMEN TEKNIK INFORMATIKA

FAKULTAS TEKNIK

UNIVERSITAS HASANUDDIN

GOWA

2024

PENGAJUAN TESIS

KONSISTENSI PENULISAN JURNAL ILMIAH BERBASIS NATURAL LANGUAGE PROCESSING

Tesis

Sebagai Salah Satu Syarat Mencapai Gelar Magister

Program Studi Magister Teknik Informatika

Disusun dan diajukan oleh

SITTI MAWADDAH UMAR

D082211023

Kepada

FAKULTAS TEKNIK
UNIVERSITAS HASANUDDIN
GOWA
2024

TESIS

KONSISTENSI PENULISAN JURNAL ILMIAH BERBASIS NATURAL LANGUAGE PROCESSING

SITTI MAWADDAH UMAR
D082211023

Telah dipertahankan di hadapan Panitia Ujian Tesis yang dibentuk dalam rangka penyelesaian studi pada Program Magister Teknik Informatika Fakultas Teknik
Universitas Hasanuddin
Pada tanggal 04 November 2024
dan dinyatakan telah memenuhi syarat kelulusan

Menyetujui,

Pembimbing Utama

Dr. Ir. Ingrid Nurtanio, M.T.
NIP. 19610813 198811 2 001

Pembimbing Pendamping

Dr. Ir. Zahir Zainuddin, M.Sc.
NIP. 19640427 198910 1 002

Dekan Fakultas Teknik
Universitas Hasanuddin

Prof. Dr.Eng. Ir. Muhammad Isran Ramli, M.T. IPM., ASEAN.Eng.
NIP. 19730926 200012 1 002

Ketua Program Studi
S2 Teknik Informatika

Dr. Ir. Zahir Zainuddin, M.Sc.
NIP. 19640427 198910 1 002

PERNYATAAN KEASLIAN TESIS DAN PELIMPAHAN HAK CIPTA

Yang bertanda tangan di bawah ini

Nama	: Sitti Mawaddah Umar
Nomor mahasiswa	: D082211023
Program studi	: Magister Teknik Informatika

Dengan ini menyatakan bahwa, tesis yang berjudul "KONSISTENSI PENULISAN JURNAL ILMIAH BERBASIS NATURAL LANGUAGE PROCESSING " adalah benar karya saya dengan arahan dari komisi pembimbing Dr. Ir. Ingrid Nurtanio dan Dr. Ir. Zahir Zainuddin, M.Sc. Karya ilmiah ini belum diajukan dan tidak sedang diajukan dalam bentuk apapun kepada perguruan tinggi manapun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka tesis ini. Sebagian dari isi tesis ini telah dipublikasikan di konferensi International (25th International Seminar on Intelligent Technology and Its Applications (ISITIA) 2024). Sebagai artikel dengan judul "Analysis of Consistency and Structure of Scholarly Papers Using Natural Language Processing".

Dengan ini saya melimpahkan hak cipta ini dari karya tulis saya berupa tesis ini kepada Universitas Hasanuddin.

Gowa, 12 November 2024

Yang menyatakan

Sitti Mawaddah Umar

KATA PENGANTAR

Alhamdulillahi rabbil alamin, segala puji bagi Allah Subhanahu Wa Taala Yang Maha Sempurna, yang telah memberikan rahmat, hidayah dan pertolongan-Nya sehingga penulis dapat menyelesaikan tesis dengan judul "konsistensi penulisan konten jurnal ilmiah berbasis natural language processing". Tak lupa pula shalawat dan salam kepada Nabi Muhammad Shalallahu Alaihi Wasallam yang telah menyinari dunia dengan keindahan ilmu dan akhlak yang diajarkan kepada seluruh umatnya.

Tesis ini disusun untuk memenuhi persyaratan untuk memperoleh gelar Magister Komputer (M.Kom) pada Program Pascasarjana Departemen Teknik Informatika. Penulis ini tidak lepas dari bantuan, bimbingan serta dukungan dari berbagai pihak, dari masa perkuliahan sampai dengan masa penyusunan tesis. Oleh karena itu, penulis dengan senang hati menyampaikan banyak terima kasih kepada:

1. Allah Subhanahu Wa Ta'ala Yang Maha Sempurna atas semua berkah, karunia, serta pertolongan-Nya yang diberikan kepada penulis disetiap Langkah dalam pembuatan program hingga penulisan laporan tesis ini.
2. Kedua Orang tua penulis, Bapak Umar Husain, S.Tr.Gz dan Ibu Maryani, S.H, yang selalu menjadi motivasi terbesar dalam penyelesaian perkuliahan ini yang tidak pernah putus memberikan doa, dukungan dan semangat serta selalu sabar dalam mendidik penulis sejak kecil.
3. Sewang, S.E.,M.M yang telah senantiasa membantu dan memberikan motivasi terhadap saya supaya selalu giat dan menuntaskan penelitian ini.
4. Ibu Dr. Ir. Ingrid Nurtanio, M.T. selaku pembimbing I dan Bapak Dr. Ir. Zahir Zainuddin, M.Sc selaku pembimbing II yang telah memberikan waktu, tenaga, pikiran, dukungan moril maupun materil serta perhatian yang luar biasa untuk mengarahkan penulis dalam pengerjaan program dan penyusunan tesis.

5. Bapak Dr. Amil Ahmad Ilham, S.T., M.IT., Bapak Prof. Dr. Ir. Ansar Suyuti, M.T., dan Ibu Muharramah Yusuf, B.Sc., M.Sc., Ph.D. selaku dosen penguji yang telah memberikan kritik dan saran yang membangun sehingga laporan tesis ini menjadi lebih baik.
6. Bapak Prof. Dr. Ir. Indrabayu, ST, MT, M.Bus.Sys., IPM, ASEAN, Eng. selaku Ketua Departemen Teknik Informatika, Fakultas Teknik Universitas Hasanuddin yang telah memberikan motivasi, bimbingan, dan dukungan selama masa perkuliahan penulis.
7. Para sahabat, saudara, dan rekan-rekan di Laboratorium Computer Based System Pascasarjana UNHAS yang telah memberikan begitu banyak bantuan, keceriaan dan dukungan selama proses perkuliahan.
8. Ibu Yuanita serta segenap Staff Departemen Magister Teknik Informatika yang telah banyak membantu penulis selama pengurusan administrasi

Penulis menyadari bahwa tesis masih jauh dari kata sempurna dan di dalam penyelesaiannya masih menemui kesulitan dan hambatan, sehingga penulis tetap mengharapkan saran dan kritik untuk pengembangan lebih lanjut, agar dapat memberikan manfaat yang banyak bagi semua pembaca.

Gowa, 17 Oktober 2024

Sitti Mawaddah Umar

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

SITTI MAWADDAH UMAR. *Analisis Konsistensi Penulisan Konten Jurnal* (dibimbing oleh **Ingrid Nurtanio** dan **Zahir Zainuddin**).

Konsistensi dalam penulisan jurnal ilmiah memiliki peran krusial dalam memandu pembaca melalui alur pemikiran dan informasi yang disajikan. Dengan memelihara konsistensi, penulis dapat menghindari kesalahan dan inkonsistensi yang berpotensi mempengaruhi hasil dan interpretasi penelitian. Penelitian ini mengembangkan sebuah sistem berbasis informasi yang bertujuan untuk mengevaluasi konsistensi dalam penulisan jurnal dengan pengujian sistem yang menghasilkan nilai coherence. Teknik-teknik Natural Language Processing (NLP), seperti Term Frequency-Inverse Document Frequency (TF-IDF) dan *Cosine Similarity* (CS), digunakan untuk menganalisis kesamaan antar bagian konten jurnal: judul, abstrak, pendahuluan, dan kesimpulan. Selain itu, sistem juga memanfaatkan Bidirectional Encoder Representations from Transformers (BERT) untuk menghasilkan embedding teks yang lebih kaya dan representatif terhadap konteks. Jarak antar embedding ini diukur menggunakan Euclidean Distance untuk mendapatkan evaluasi yang lebih mendalam mengenai koherensi dan keselarasan antar bagian teks. Dengan mengombinasikan perhitungan *Cosine Similarity* dan Euclidean Distance melalui embedding BERT.

Kata Kunci: *Natural Language Processing (NLP), Term Frequency-Inverse Document Frequency (TF-IDF), Bidirectional Encoder Representations from Transformers (BERT), Euclidean Distance, Text Summarization, Cosine Similarity*

ABSTRACT

SITTI MAWADDAH UMAR. *Consistency Analysis of Journal Content Writing.* (Supervised by Ingrid Nurtanio, and Zahir Zainuddin).

Consistency in scientific journal writing is crucial in guiding readers through the flow of thought and the information presented. By maintaining consistency, authors can avoid errors and inconsistencies affecting the research results and interpretation. This study develops an information-based system to evaluate the consistency in journal writing through system testing that yields coherence scores. Natural Language Processing (NLP) techniques, such as Term Frequency-Inverse Document Frequency (TF-IDF) and *Cosine Similarity* (CS), are employed to analyze the similarity between different sections of the journal content: title, abstract, introduction, and conclusion. Additionally, the system leverages Bidirectional Encoder Representations from Transformers (BERT) to generate more prosperous and more contextually representative text embeddings. The distance between these embeddings is calculated using Euclidean distance to evaluate the coherence and alignment between different text sections in more depth. By combining *Cosine Similarity* and Euclidean Distance through BERT embeddings, the system provides optimized results in assessing the overall consistency of scientific journal writing.

Keywords: Natural Language Processing (NLP), Term Frequency-Inverse Document Frequency (TF-IDF), Bidirectional Encoder Representations from Transformers (BERT), Euclidean Distance, Text Summarization, *Cosine Similarity*

BAB I

PENDAHULUAN

1.1 Latar Belakang

Penulisan jurnal atau karya tulis ilmiah, merupakan karya tulis ilmiah yang dijadikan peneliti sebagai media dalam menyalurkan atau menjelaskan kepada pembaca mengenai penelitian yang tengah atau selesai dijalankan. Penyusunan jurnal tersebut memerlukan pedoman konsistensi penulisan yang harus di perhatikan antar tiap konten jurnal yang tersedia serta analisis yang objektif dan menggunakan data yang empiris (Wang & Liu, 2017). Topik – topik yang terdapat di jurnal dapat menjadi landasan ilmu pengembangan atau terciptanya sebuah inovasi baru dalam penelitian selanjutnya, namun pembuatan karya ilmiah atau jurnal tidak dapat disusun secara bebas.

Konsistensi penulisan dalam karya tulis ilmiah dijadikan sebagai syarat dalam penerbitan termasuk konsistensi penulis dalam menulis karya ilmiah dengan memperhatikan keterkaitan antara tiap konten yang tertera dalam format penulisan karya ilmiah sesuai dengan pedoman penulisan yang disediakan oleh penyedia layanan publikasi. Proses penelitian tersebut membuat para penulis naskah jurnal diharuskan untuk mampu menyesuaikan dan menyusun naskahnya secara konsisten dalam penulisan dengan topik penelitian yang diusulkan, serta isi konten jurnal yang disusun oleh penulis. Hal tersebut memberikan tantangan pada penulis untuk mampu melakukan analisis mandiri dalam jurnal yang disusunnya untuk dapat memenuhi tuntutan dari penyedia layanan (Prominski et al., 2018).

Ditinjau dari perkembangan penulisan ilmiah, menjadi jelas bahwa tidak semua naskah jurnal memiliki tingkat konsistensi dan struktur yang optimal. Konsistensi yang baik dalam penyajian gagasan dan struktur yang terorganisir dengan baik dapat meningkatkan daya jangkau dan pemahaman pembaca. Oleh karena itu, perlu ada perhatian lebih terhadap evaluasi dan analisis konsistensi serta struktur dalam naskah jurnal yang dimuat dalam sebuah sistem.

Dalam hal ini, seberapa konsisten penulis dalam menyusun naskahnya memerlukan gaya penulisan yang baik dengan penggunaan tata bahasa yang lebih

mudah dipahami, serta penggunaan istilah dan konvensi tertentu dalam disiplin ilmu di bidangnya masing-masing. Tantangan dalam penelitian ini adalah bagaimana penulis jurnal mampu menyusun sebuah jurnal dengan model penulisan yang konsisten, bukan hanya dituntut untuk konsisten dalam menyelesaikan tulisannya namun juga mampu membuat pembaca mengetahui apa yang masalah dan objek penelitian yang sedang diteliti. Sistem ini diproses sehingga menghasilkan nilai konsisten dari penulis dan dikemas menjadi sebuah sistem informasi. Terkait konsistensi penulisan tersebut kami menentukan parameter kesesuaian antar bagian konten jurnal dengan masing-masing jumlah kata yang terdapat di Judul, Abstrak, Pendahuluan, dan kesimpulan yang akan disajikan dalam sebuah jurnal. Konten atau topik jurnal tersebut membutuhkan keterkaitan atau relevansi antara masing-masing kontennya sebelum dipublikasikan.

Sistem yang akan dibangun merupakan sebuah sistem yang menggunakan bahasa pemrograman *Phyton*, dengan menggunakan desain *web responsive* (Rinartha & Suryasa, 2017). Penelitian ini diharapkan mampu menguji konsistensi penulis dalam menyusun makalah. Untuk meningkatkan relasi antara judul, abstrak, pendahuluan, dan kesimpulan dengan menggunakan sistem pengklasifikasian yang menghasilkan informasi akurat tentang kesesuaian jurnal yang disusun penulis dalam mengolah kata dengan gaya penulisan, gaya kutipan(Zhao et al., 2020), serta penggunaan kalimat aktif atau pasif yang lebih mudah dipahami pembaca.

Dalam tahapan tersebut proses analisis teks mendalam menggunakan beberapa tahapan proses yang umum yang dilakukan dalam *Natural Language Processing* (NLP), yaitu: Tokenisasi, *Stopword removal*, *parsing*, *Stemming.* Proses tersebut mengacu pada susunan kata-kata dalam sebuah kalimat sehingga tampak masuk akal secara tata bahasa (Alexandrov et al., 2016). Teks yang telah diolah dalam tahapan NLP menghasilkan sebuah kalimat yang berupa simpulan sederhana seperti sistem inforrmasi. Dengan menambahkan algortitma *Bidirectional Encoder Representations from Transformers* (BERT) yang digunakan untuk memahami pemaknaan kalimat dalam model bahasa yang dilatih

menggunakan arsitektur *transformer* yang memungkinkan pemahaman kontekstual dari kata-kata dalam sebuah kalimat(Ramina et al., 2020).

Sedangkan untuk algoritma *Cosine Similarity* dapat digunakan membandingkan kata-kata artikel dengan subtopik/deskripsi dalam topik penelitian (Nursalman et al., n.d.). Proses tersebut dapat lebih efisien dalam sebuah kinerja sistem yang dibangun dapat berjalan dengan baik dan stabil dengan minimnya kegagalan dan kesalahan dalam penilaian, maka diusulkan TF-IDF (*Term Frequency - Inverse Document Frequency)* dalam menentukan bobot dari setiap kata dalam sebuah dokumen (Zaware et al., 2021). Pengembangan sistem secara keseluruhan dipengaruhi oleh setiap masing-masing kata yang mendeteksi konsistensi penulis dalam membuat makalah yang relevan.

Ekstraksi bobot dilakukan untuk menghitung nilai suatu perangkat lunak yang menghasilkan *coherence score* yang akurat, dengan hitungannya berbasis *scientific* dan relevan(Krishnaveni & Balasundaram, 2017). Nilai *coherence* atau nilai hasil dari *Cosine Similarity* merupakan nilai konsistensi yang didapatkan dari pengujian sistem. Sistem kesamaan kata yang dibangun dapat menjadi acuan untuk hasil yang akurat, dengan mengacu pada format IEEE *(Institute of Electrical and Electronics Engineers)* dan ACM *(Association for Computing Machinery)*(Ghosal et al., 2019). Metode penyelesaian ini diharapkan dapat menghasilkan *output* yang relevan dalam penulisan naskah jurnal yang belum diterbitkan maupun jrunal yang telah diterbitkan, untuk mengetahui konsistensi penulis dalam menyusun makalah berdasarkan judul, abstrak, pendahuluan, dan kesimpulan serta mampu menghasilkan sebuah simpulan sederhana, memberikan makna yang mudah dipahami oleh pembaca.

1.2 Rumusan Masalah

Berdasarkan deskripsi latar belakang masalah, rumusan masalah pada penelitian ini adalah Bagaimana sistem berbasis algoritma NLP, TF-IDF, BERT, dan *Cosine Similarity* dapat mendeteksi dan mengevaluasi konsistensi konten dalam jurnal ilmiah melalui analisis kemiripan teks antarbagian (*Title, Abstract, Introduction, Conclusion*)?

1.3 Tujuan Penelitian

Penelitian ini bertujuan untuk menguji sistem berbasis algoritma NLP, TF-IDF, BERT, dan *Cosine Similarity* yang mampu mendeteksi serta mengevaluasi konsistensi konten dalam jurnal ilmiah melalui analisis kemiripan teks antarbagian, khususnya pada Judul, Abstrak, Pendahuluan, dan Kesimpulan. Sistem ini diharapkan dapat memberikan wawasan tentang keselarasan konteks dan koherensi penulisan antarbagian dalam jurnal ilmiah.

1.4 Manfaat Penelitian

Berdasarkan Membantu untuk mengetahui seberapa konsisten penulis dalam menulis karya ilmiah dengan mendeteksi kecocokan teks dan inkonsistensi penulisan jurnal akademik yang berdasar pada judul, abstrak, pendahuluan, dan kesimpulan.

1.5 Batasan Masalah

Adapun Batasan masalah dalam penelitian ini yaitu:

1. Data yang digunakan dibatasi pada bagian judul, abstrak, pendahuluan, dan kesimpulan dari setiap artikel jurnal.
2. Analisis konsistensi penulisan hanya akan dilakukan dengan menggunakan metode *Natural Language Processing* (NLP), *Term Frequency-Inverse Document Frequency* (TF-IDF), *Cosine Similarity*, *Euclidean Distance* dan *Bidirectional Encoder Representations from Transformers* (BERT).
3. Evaluasi konsistensi penulisan akan diukur dengan nilai similaritas atau *coherence* antara bagian yang relevan.
4. Simpulan sederhana konten jurnal hanya akan menggunakan kemampuan BERT untuk menghasilkan kesimpulan yang mudah dipahami.
5. Pengoptimalan yang dilakukan menggunakan perpaduan dari BERT dan *Eulidean Distance*.

BAB II

TINJAUAN PUSTAKA

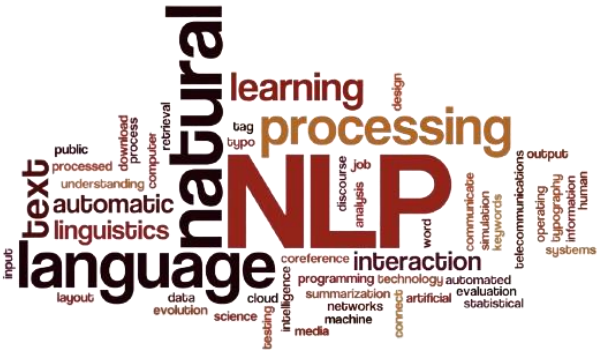
1.1 Kajian Pustaka

Kajian pustaka yang tertuang pada bab ini hasil dari studi pendahuluan yang telah dilaksanakan, studi pendahuluan yang dilakukan adalah studi literatur dengan melaksanakan review terhadap jurnal internasional yang relevan dengan tema penelitian, mereview buku, diskusi dan modul yang mendukung materi, melaksanakan browsing di internet dan juga menganalisis video yang relevan.

1.2 Landasan Teori

1.2.1 *Natural Language Processing* (NLP)

Natural Language Processing (NLP), juga dikenal sebagai komputasi linguistik, mengkonsolidasikan dirinya sebagai bidang penelitian yang melibatkan model komputasi dan proses untuk memecahkan masalah praktis untuk memahami dan memanipulasi bahasa manusia. Terlepas dari bentuk manifestasinya, tekstual atau ucapan, bahasa alami dipahami sebagai segala bentuk komunikasi sehari-hari antara manusia. Definisi ini mengecualikan bahasa pemrograman dan notasi matematika, yang dianggap sebagai bahasa buatan. Bahasa alami terus berubah, sehingga sulit untuk menetapkan aturan eksplisit untuk komputer (de Oliveira et al., 2021).

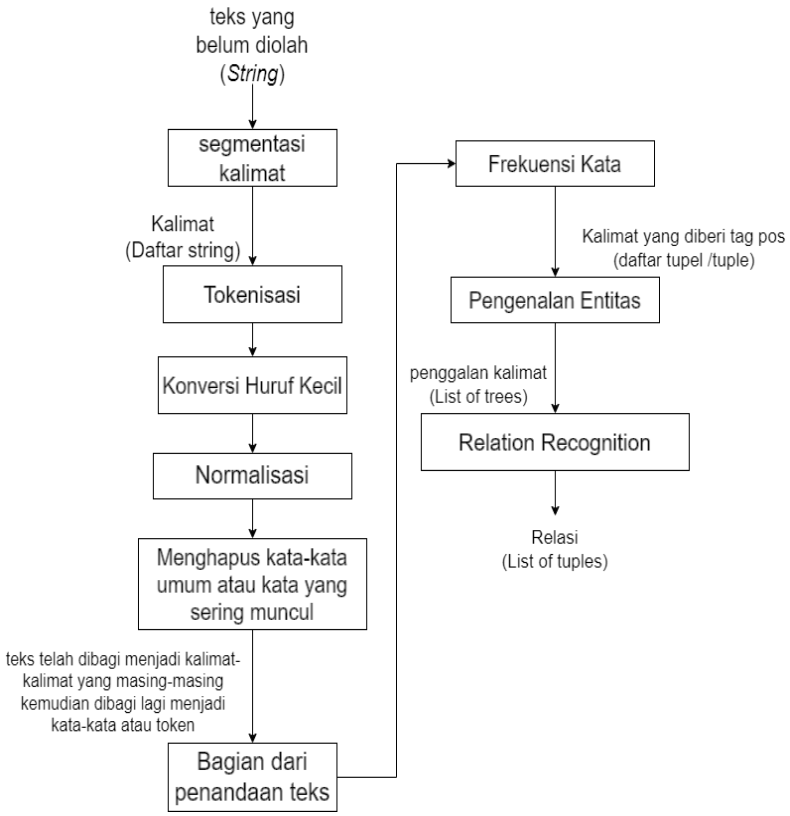


Gambar 1. Ilustrasi Pemrosesan Teks

Natural Language Processing (NLP) adalah cabang dari kecerdasan buatan (*artificial intelligence*) yang berfokus pada interaksi antara komputer dan

bahasa manusia alami baik dalam bentuk teks ataupun s*peech*. Tujuan utama NLP adalah memberikan kemampuan komputer untuk memahami, menafsirkan, dan menghasilkan bahasa manusia dengan cara yang bermakna. Dengan kata lain, NLP mencoba untuk memungkinkan komunikasi yang efektif antara manusia dan mesin melalui bahasa alami. (Sharma et al., 2021)

Oleh karena itu, pemrosesan bahasa natural (*Natural Language Processing*) dibutuhkan dalam setiap proses otomatisasi yang melibatkan dokumen teks seperti halnya dalam penelitian ini. Penggunaan NLP dalam suatu sistem bukan berarti kemampuan untuk memahami teks secara sepenuhnya, tetapi lebih bertujuan untuk melakukan ekstraksi konsep yang terdapat pada suatu dokumen. Contoh alur pemrosesan teks dengan NLP bisa dilihat pada gambar 2.



Gambar 2. Arsitektur sistem *Pipeline* ekstraksi informasi berbasis dokumen teks

Dalam *Natural Language Processing* (NLP) terkadang terjadi ambiguitas dalam membaca maupun memahami sebuah data, cara mengatasi ambiguitas ini sering kali melibatkan model pembelajaran mesin yang dilatih pada korpus besar

teks untuk memahami konteks yang lebih luas (untuk *context-dependent ambiguity)* atau menggunakan kamus dan data leksikal untuk menjelaskan makna yang berbeda (untuk *context-independent ambiguity*), (Seong et al., 2023).

Dalam konteks penelitian ini, penggunaan *natural language processing* (NLP) lebih ditekankan pada ekstraksi konteks suatu kalimat dalam dokumen dan suatu kata dalam kalimat (Kilany et al., 2018), karena fokus utamanya adalah mekanisme penggalian kalimat kebutuhan dan kata kunci yang merupakan representasi dari fitur pada suatu dokumen. Proses yang akan dilakukan pertama adalah teks dokumen akan diekstraksi menjadi kalimat, kemudian kalimat-kalimat tersebut akan dibagi lagi menjadi kata-kata dengan penomoran tertentu, yang disebut dengan tokenisasi(Amanullah et al., 2019)

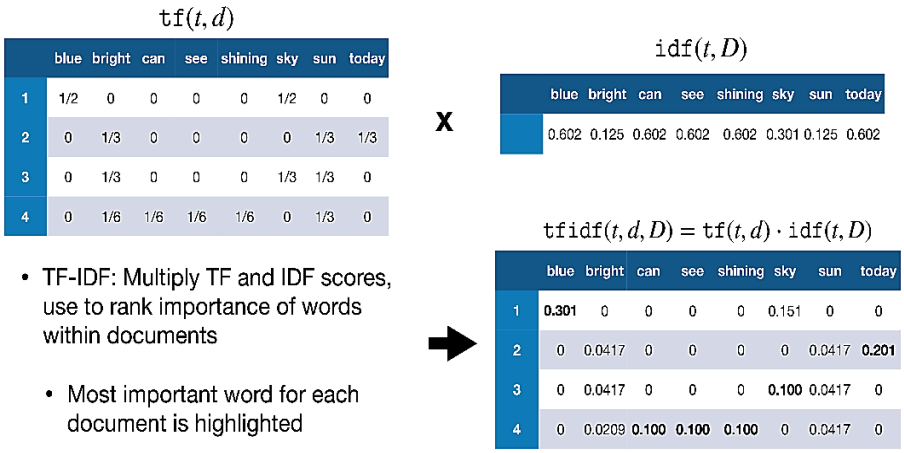
Pada dasarnya, *Natural Language Processing* (NLP) mencakup beberapa aspek utama:

1. Pemahaman Bahasa**:** NLP berusaha memungkinkan komputer untuk memahami struktur tata bahasa, sintaksis, dan bahasa manusia.
2. Pemrosesan Teks: Salah satu fokus utama NLP adalah pemrosesan teks, termasuk tugas seperti pengenalan entitas, ekstraksi informasi, dan pemahaman konteks dari teks.
3. Penghasilan Teks: NLP juga mencakup kemampuan komputer untuk menghasilkan teks secara alami, seperti dalam pembangkitan teks, penulis otomatis, dan penerjemahan mesin.
4. Pemahaman Makna: NLP berusaha memahami makna di balik kata-kata, menangani ambiguitas, dan memahami perbedaan makna dalam konteks tertentu.
5. Analisis Sentimen: Digunakan untuk menganalisis sentimen di balik teks, membantu menilai opini dan perasaan yang terkandung dalam suatu teks.
6. Penerjemahan Mesin: NLP digunakan dalam sistem penerjemahan mesin untuk menerjemahkan teks dari satu bahasa ke bahasa lainnya secara otomatis.

7. Model Bahasa dan Pembelajaran Mesin: NLP melibatkan pengembangan model bahasa dan teknik pembelajaran mesin, seperti *Word Embeddings* dan *Transformer Architecture*.
8. Tantangan: NLP dihadapkan pada berbagai tantangan, termasuk ambiguitas bahasa, variasi gaya bahasa, dan pemahaman konteks yang kompleks.

1.2.2 *Term Frequnecy - Inverse Document Frequency* (TF-IDF)

TF-IDF, algoritma adalah metode statistik tertimbang yang biasa digunakan dalam pencarian informasi dan penambangan data. Ini digunakan untuk mengevaluasi pentingnya kata-kata untuk teks atau korpus. Pentingnya kata-kata meningkat secara proporsional dengan frekuensi kemunculannya dalam teks, tetapi juga menurun secara terbalik dengan frekuensi kemunculannya dalam korpus (Raza et al., 2021). TF merupakan kependekan dari *Term Frequency*, menjelaskan frekuensi kemunculan sebuah kata dalam teks. Untuk mencegah kata yang sama menjadi lebih penting dalam teks panjang daripada teks pendek (kata yang sama memiliki frekuensi lebih tinggi dalam teks panjang daripada teks pendek, terlepas dari apakah kata itu penting atau tidak), nilainya biasanya dinormalisasi (Mz et al., 2023).



Gambar 3. Pencarian teks menggunakan TF-IDF

IDF merupakan singkatan dari *Inverse Document Frequency*, mengukur signifikansi umum dari sebuah kata. Ide sentralnya adalah bahwa kata yang

muncul di semua teks tidak dapat mewakili isi utama dari sebuah teks. Namun jika teks lain yang mengandung kata ini lebih sedikit, hal ini menunjukkan bahwa kata tersebut memiliki kemampuan klasifikasi yang baik (Lahitani et al., 2016). Oleh karena itu, ide utama dari algoritma TF-IDF adalah jika sebuah kata atau frase lebih sering muncul dalam sebuah artikel (nilai TF tinggi) dan jarang muncul di artikel lain (nilai DF rendah, nilai IDF tinggi), dianggap bahwa kata atau frasa tersebut dapat mewakili artikel dengan baik dan dapat digunakan untuk klasifikasi (Yao et al., 2019).

Di TF-IDF, nilai TF sebuah kata dihitung sebagai :

$$\boldsymbol{tf_{ij} = n\frac{n_{ij}}{\Sigma_k n_{kj}}} \quad (1)$$

Dalam rumus (1), n_{ij} menunjukkan frekuensi terjadinya t_i dalam teks d_j, dan penyebut adalah jumlah frekuensi kemunculan semua kata dalam teks d_j. Dan nilai IDF t_i, membagi jumlah total teks dengan jumlah teks yang mengandung kata t_i, dan kemudian mendapatkan hasil bagi logaritma.

$$\boldsymbol{idf_i = \frac{log|D|}{|\{j:t_i \in d_j\}|} + 1} \quad (2)$$

Dalam rumus (2) | D | mewakili jumlah total semua teks, | {j: t_i |mewakili jumlah teks yang mengandung kata t_i dalam teks (jumlah teks yang mengandung kata $n_{ij} \neq 0$). Mengingat jika kata t_i tidak ada dalam korpus, maka akan menyebabkan pembagi menjadi nol, sehingga pada umumnya penyebutnya akan +1.

$$\boldsymbol{TF - IDF_{ij} = tf_{ij} \times idf_i} \quad (3)$$

Dalam rumus (3) Nilai ini adalah hasil dari perkalian antara TF dan IDF. Memberikan bobot yang seimbang antara frekuensi kata dalam suatu dokumen dan jarangnya kata tersebut dalam seluruh koleksi dokumen.

Implementasi Praktis atau Langkah-langkah:

1. Tokenisasi: Memecah dokumen menjadi kata-kata atau token.
2. Perhitungan TF: Menghitung frekuensi kemunculan setiap kata dalam dokumen.

3. Perhitungan IDF: Menghitung nilai IDF untuk setiap kata dalam koleksi dokumen.
4. Perhitungan TF-IDF: Mengalikan nilai TF dengan nilai IDF untuk setiap kata dalam setiap dokumen.

Berikut adalah contoh perhitungan TF-IDF pada suatu kalimat yang telah dilakukan *pre-processing* pada penelitian (Khusna & Agustina, 2018), misalnya:

D1(Q) = "sistem pakar sistem informasi isi tahu pakar guna konsultasi"

D2 = "cerdas buat mampu sistem tafsir data ajar data sebut guna capai tuju tugas tentu"

D3 = "sistem pakar suatu program komputer rancang model mampu selesai masalah laku orang pakar"

D4 = "sistem dukung putus bantu pecah masalah tentu produk baik ideal"

Tabel 1. Hasil *Pre-processing*

[Kata], [ajar], [baik], [Bantu], [buat], [capai], [cerdas], [data], [dukung], [guna], [ideal], [informasi], [isi], [komputer], [konsultasi], [laku], [mampu], [masalah], [model], [orang], [pakar], [pecah], [produk], [program], [putus], [rancang], [sebut], [selesai], [sistem], [suatu], [tafsir], [tahu], [tentu], [tugas], [tuju].

Tabel 2. Perhitungan *Term Frequency* (TF)

***Term*/kata**	**TF**			
	D1	**D2**	**D3**	**D4**
ajar	0	1	0	0
baik	0	0	0	1
bantu	0	0	0	1
buat	0	1	0	0
capai	0	1	0	0
cerdas	0	1	0	0
data	0	2	0	0
dukung	0	0	0	1
guna	1	1	0	0
ideal	0	0	0	1
informasi	1	0	0	0
isi	1	0	0	0
komputer	0	0	1	0

***Term*/kata**	**TF**			
	D1	**D2**	**D3**	**D4**
konsultasi	1	0	0	0
laku	0	0	1	0
mampu	0	1	1	0
masalah	0	0	1	1
model	0	0	1	0
orang	0	0	1	0
pakar	2	0	2	0
pecah	0	0	0	1
produk	0	0	0	1
program	0	0	1	0
putus	0	0	0	1
rancang	0	0	1	0
sebut	0	1	0	0
selesai	0	0	1	0
sistem	2	1	1	1
suatu	0	0	1	0
tafsir	0	1	0	0
tahu	1	0	0	0
tentu	0	1	0	1
tugas	0	1	0	0
tuju	0	1	0	0

Tabel 2 menampilkan nilai TF dari masing-masing *term*/kata, di mana TF merupakan banyaknya *term*/kata yang muncul dalam suatu dokumen. Dapat diketahui bahwa kata "sistem" muncul sebanyak 2 kali pada dokumen 1, 1 kali pada dokumen 2, 1 kali pada dokumen 3, dan 1 kali pada dokumen 4. Begitu pula perhitungan TF untuk kata lainnya dalam dokumen. Tahapan selanjutnya adalah menghitung nilai IDF.

Tabel 3. Perhitungan IDF

***Term*/kata**	**DF**	**IDF** $\log((D/df+1))$
ajar	1	$\log(4/1)+1=1{,}602$
baik	1	$\log(4/1)+1=1{,}602$
bantu	1	$\log(4/1)+1=1{,}602$
buat	1	$\log(4/1)+1=1{,}602$
capai	1	$\log(4/1)+1=1{,}602$

***Term*/kata**	**DF**	**IDF** $\log((D/df+1))$
cerdas	1	$\log(4/1)+1=1{,}602$
data	1	$\log(4/1)+1=1{,}602$
dukung	1	$\log(4/1)+1=1{,}602$
guna	2	$\log(4/2)+1=1{,}301$
ideal	1	$\log(4/1)+1=1{,}602$
informasi	1	$\log(4/1)+1=1{,}602$
isi	1	$\log(4/1)+1=1{,}602$
komputer	1	$\log(4/1)+1=1{,}602$
konsultasi	1	$\log(4/1)+1=1{,}602$
laku	1	$\log(4/1)+1=1{,}602$
mampu	2	$\log(4/2)+1=1{,}301$
masalah	2	$\log(4/2)+1=1{,}301$
model	1	$\log(4/1)+1=1{,}602$
orang	1	$\log(4/1)+1=1{,}602$
pakar	2	$\log(4/2)+1=1{,}301$
pecah	1	$\log(4/1)+1=1{,}602$
produk	1	$\log(4/1)+1=1{,}602$
program	1	$\log(4/1)+1=1{,}602$
putus	1	$\log(4/1)+1=1{,}602$
rancang	1	$\log(4/1)+1=1{,}602$
sebut	1	$\log(4/1)+1=1{,}602$
selesai	1	$\log(4/1)+1=1{,}602$
sistem	4	$\log(4/4)+1=1$
suatu	1	$\log(4/1)+1=1{,}602$
tafsir	1	$\log(4/1)+1=1{,}602$
tahu	1	$\log(4/1)+1=1{,}602$
tentu	2	$\log(4/2)+1=1{,}301$
tugas	1	$\log(4/1)+1=1{,}602$
tuju	1	$\log(4/1)+1=1{,}602$

Pada perhitungan IDF, sebelumnya perlu diketahui terlebih dahulu nilai *D* dan DF nya, di mana *D* merupakan jumlah semua dokumen yang ada pada *dataset*, sedangkan DF merupakan jumlah dokumen yang mengandung *term (t)*. Berdasarkan Tabel 3, misalnya pada kata "sistem", diketahui bahwa nilai DF-nya adalah 4, yang berarti bahwa kata tersebut muncul pada 4 dokumen. Sehingga didapatkan nilai IDF untuk kata "sistem" adalah 1. Sebagai contoh:

$$IDF = \log(n/df) + 1 \qquad (4)$$

Dimana n adalah banyaknya dokumen. Untuk kata “sistem” dan “guna”:

$$IDF(sistem) = log\ (4/4) + 1 = 1$$

$$IDF(guna) = \log (4/2) + 1 = 1{,}301$$

Setelah didapatkan nilai TF dan IDF, selanjutnya akan dihitung nilai TF-IDF atau nilai bobot dengan mengalikan nilai TF dengan nilai IDF.

Tabel 4. Perhitungan TF-IDF

Term/kata	TF				DF	IDF	BOBOT(W)=TF*IDF			
	D1	D2	D3	D4			D1	D2	D3	D4
ajar	0	1	0	0	1	1,602	0	1,602	0	0
baik	0	0	0	1	1	1,602	0	0	0	1,602
bantu	0	0	0	1	1	1,602	0	0	0	1,602
buat	0	1	0	0	1	1,602	0	1,602	0	0
capai	0	1	0	0	1	1,602	0	1,602	0	0
cerdas	0	1	0	0	1	1,602	0	1,602	0	0
data	0	2	0	0	1	1,602	0	3,204	0	0
dukung	0	0	0	1	1	1,602	0	0	0	1,602
guna	1	1	0	0	2	1,301	1,301	1,301	0	0
ideal	0	0	0	1	1	1,602	0	0	0	1,602
informasi	1	0	0	0	1	1,602	1,602	0	0	0
isi	1	0	0	0	1	1,602	1,602	0	0	0
komputer	0	0	1	0	1	1,602	0	0	1,602	0
konsultasi	1	0	0	0	1	1,602	1,602	0	0	0
laku	0	0	1	0	1	1,602	0	0	1,602	0
mampu	0	1	1	0	2	1,301	0	1,301	1,301	0
masalah	0	0	1	1	2	1,301	0	0	1,301	1,301
model	0	0	1	0	1	1,602	0	0	1,602	0
orang	0	0	1	0	1	1,602	0	0	1,602	0
pakar	2	0	2	0	2	1,301	2,602	0	2,602	0
pecah	0	0	0	1	1	1,602	0	0	0	1,602
produk	0	0	0	1	1	1,602	0	0	0	1,602
program	0	0	1	0	1	1,602	0	0	1,602	0
putus	0	0	0	1	1	1,602	0	0	0	1,602
rancang	0	0	1	0	1	1,602	0	0	1,602	0
sebut	0	1	0	0	1	1,602	0	1,602	0	0
selesai	0	0	1	0	1	1,602	0	0	1,602	0
sistem	2	1	1	1	4	1	2	1	1	1
suatu	0	0	1	0	1	1,602	0	0	1,602	0
tafsir	0	1	0	0	1	1,602	0	1,602	0	0
tahu	1	0	0	0	1	1,602	1,602	0	0	0
tentu	0	1	0	1	2	1,301	0	1,301	0	1,301
tugas	0	1	0	0	1	1,602	0	1,602	0	0
tuju	0	1	0	0	1	1,602	0	1,602	0	0

Berdasarkan Tabel 4, didapatkan bahwa pada kata “pakar” pada dokumen 1 memiliki nilai TF yaitu 2, dokumen 3 memiliki nilai TF yaitu 2, dan nilai IDF yaitu 1,301. Sehingga hasil perhitungan nilai bobot atau TF-IDF kata tersebut pada dokumen 1 dan 3 adalah 2,602.

TF-IDF("pakar" ,D1)=2*1.301=2,602

1.2.3 Algoritma *Cosine Similarity*

Cosine Similarity berfungsi untuk membandingkan kemiripan antar dokumen, dalam hal ini yang dibandingkan adalah *query* dengan dokumen latih (Pattnaik & Nayak, 2019). Dalam menghitung *Cosine Similarity*, pertama yang di lakukan yaitu melakukan perkalian skalar antara query dengan dokumen kemudian dijumlahkan, setelah itu melakukan perkalian antara panjang dokumen dengan panjang *query* yang telah dikuadratkan, setelah itu di hitung akar pangkat dua. Selanjutnya hasil perkalian skalar tersebut di bagi dengan hasil perkalian panjang dokumen dan *query* yang mana pada pengujian sistem ini menghasilkan nilai similaritas yang sama dengan nilai *coherence* yaitu 0 dan 1 yang diartikan apabila nilai 0 mendekati angka 1 maka hasil yang didapatkan akan semakin baik.

Cosine Similarity juga merupakan metode pengukuran kesamaan antara dua vektor dalam ruang berdimensi banyak, terutama digunakan dalam pemrosesan teks dan pengambilan informasi. Metode ini mengukur sejauh mana dua vektor mendekati arah yang sama.

Rumus dapat dilihat sebagai berikut :

$$\boldsymbol{cosSim}(\boldsymbol{d_j}, \boldsymbol{q_k}) = \frac{\sum_{i=1}^{n}(\boldsymbol{tdij} \times \boldsymbol{tiqk})}{\sqrt{\sum_{i=1}^{n} \boldsymbol{tdij^2}} \times \sum_{t=1}^{n} \boldsymbol{tqik^2}} \qquad (5)$$

Keterangan :

cosSim : tingkat kesamaan dokumen dengan query tertentu
tdij : term ke-i dalam vektor untuk dokumen ke-j
tqik : term ke-i dalam vektor untuk query ke-k
n : jumlah *term* yang unik dalam data set

Langkah-langkah perhitungan manual algoritma *Cosine Similarity*.

1. Ditentukan terlebih dahulu masing-masing *query*, yaitu *query* dari jawaban (D), *query* dari *key* jawaban (Q) dan gabungan keduanya *(Queries*). Untuk meencari nilai dari DxQ yaitu menggunakan perkalian skalar TF-IDF yang masing-masing D terhadap TF-IDF Q, kemudian dicari totalnya. Misalnya:

$$DxQ\ (\text{"guna"}, D1\ thd\ Q) = TF - IDF(D1, \text{"guna"}) * TF - IDF(Q"guna") = 1{,}301 * 1{,}301 = 1{,}692$$

$$DxQ\ (\text{"Pakar"}, D2\ thd\ Q) = TF - IDF(D2, \text{"Pakar"}) * TF - IDF(Q"Pakar") = 2{,}602 * 2{,}602 = 6{,}770$$

$$DxQ\ (\text{"Baik"}, D3\ thd\ Q) = TF - IDF(D3, \text{"Baik"}) * TF - IDF(Q\text{"Baik"}) = 1{,}602 * 0 = 0$$

2. Mencari kuadrat TF-IDF. Kemudian dicari totalnya dan diakar kuadrat :

$$Kuadrat\ TF - IDF(\text{"Guna"}, D1) = 1{,}301\sqrt{2} = 1{,}692$$

$$Kuadrat\ TF - IDF(\text{"Data"}, D2) = 3{,}204\sqrt{2} = 10{,}266$$

$$Kuadrat\ TF - IDF(\text{"Ajar"}, D3) = 0\sqrt{2} = 0$$

Tabel 5. Hasil akar kuadrat TF-IDF

Kata/*Term*	**D1**	**D2**	**D3**	**D4**
ajar	0	2,566,596,216	0	0
baik	0	0	0	2,566,596,216
bantu	0	0	0	2,566,596,216
buat	0	2,566,596,216	0	0
capai	0	2,566,596,216	0	0
cerdas	0	2,566,596,216	0	0
data	0	1,026,638,486	0	0
dukung	0	0	0	2,566,596,216
guna	169,267,905	169,267,905	0	0
ideal	0	0	0	2,566,596,216
informasi	2,566,596,216	0	0	0
isi	2,566,596,216	0	0	0
komputer	0	0	2,566,596,216	0
konsultasi	2,566,596,216	0	0	0
laku	0	0	2,566,596,216	0
mampu	0	169,267,905	169,267,905	0
masalah	0	0	169,267,905	169,267,905
model	0	0	2,566,596,216	0
orang	0	0	2,566,596,216	0

Kata/*Term*	D1	D2	D3	D4
pakar	6,770,716,198	0	6,770,716,198	0
pecah	0	0	0	2,566,596,216
produk	0	0	0	2,566,596,216
program	0	0	2,566,596,216	0
putus	0	0	0	2,566,596,216
rancang	0	0	2,566,596,216	0
sebut	0	2,566,596,216	0	0
selesai	0	0	2,566,596,216	0
sistem	4	1	1	1
suatu	0	0	2,566,596,216	0
tafsir	0	2,566,596,216	0	0
tahu	2,566,596,216	0	0	0
tentu	0	169,267,905	0	169,267,905
tugas	0	2,566,596,216	0	0
tuju	0	2,566,596,216	0	0
Total Kuadrat	2,272,978,011	3,687,719,174	3,168,884,402	2,235,153,161
Akar Kuadrat	4,767,575,916	6,072,659,363	5,629,284,504	4,727,740,645

3. Kemudian mecari hasil *Cosine Similarity* tiap D dari [total DxQ] tiap D dibagi perkalian antara [total akar kuadrat dari kuadrat TF-IDF] milik masing-masing D dan [total aka kuadrat dari kuadrta TF-IDF] milik Q.

$$Cosine\ Similarity\ (D1) = 3{,}9267905/(4{,}76757916 * 6{,}072659363)$$

$$Cosine\ Similarity\ (D2) = 8{,}770716198/(4{,}76757916 * 5{,}629284504)$$

$$Cosine\ Similarity\ (D3) = 2/(4{,}76757916 * 4{,}76757916)$$

Tabel 6. Hasil *Cosine Similarity*

	D2	D3	D4
Cosine Similarity	0,127545466	0,326801652	0,088731686
Persentase Kemiripan	12,75454659	32,68016516	8,873168617

Nilai kemiripan dokumen D1(*Query)* terhadap semua dokumen pembanding lainnya, sehingga ditemukan dokumen yang paling mirip terletak pada dokumen D3 dengan kemiripan 32,68016516 atau 32,68%.

2.2.4 *Machine Learning*

Machine learning (ML) adalah cabang ilmu yang berkembang dalam domain kecerdasan buatan (AI), yang memfokuskan pada pembelajaran pola dan teori komputasi. *Machine Learning* (ML) melibatkan pengembangan algoritma yang dapat belajar dari data dan membuat prediksi berdasarkan pembelajaran tersebut, daripada mengikuti instruksi program statis. Ada dua jenis tugas dalam *Machine Learning* (ML), yaitu: (Jin & Xu, 2019)

1. *Supervised machine learning*: Ini melibatkan penggunaan algoritma yang telah dilatih sebelumnya dengan data latih. Algoritma ini kemudian dapat menghasilkan prediksi yang akurat saat diberikan data baru, karena kemampuannya untuk mempelajari pola dari data latih tersebut.
2. *Unsupervised machine learning*: Dalam tugas ini, algoritma belajar untuk mendeteksi pola berdasarkan kemiripan di antara data tanpa memerlukan data latih yang telah ditentukan sebelumnya. Ini memungkinkan model untuk menemukan struktur atau pola yang mungkin tersembunyi dalam data.

2.2.5 *Bidirectional Encoder Representations from Transformers* (BERT)

Salah satu tonggak penting dalam perkembangan pemrosesan bahasa alami atau *Natural Language Processing* (NLP) yang dikembangkan oleh peneliti di *Google AI Language*, BERT mengusung arsitektur *Transformer* yang telah terbukti efektif dalam memproses teks secara paralel dan efisien. Keunggulan utama BERT dibandingkan dengan model-model sebelumnya adalah kemampuannya dalam memahami konteks dua arah dalam teks, yang memungkinkannya untuk memperoleh representasi yang lebih baik dari teks yang dianalisis. Selama proses *pre-training*, BERT dilatih pada korpus teks yang sangat besar, termasuk Wikipedia dan *BookCorpus*, menggunakan metode *unsupervised learning*.

Teknik yang digunakan dalam *pre-training* adalah *Masked Language Model* (MLM), di mana beberapa kata dalam teks diacak dan model diminta untuk memprediksi kata-kata yang diacak tersebut berdasarkan konteksnya. Setelah dilatih, BERT dapat disesuaikan *(fine-tuned)* untuk tugas-tugas khusus dalam NLP, seperti klasifikasi teks, analisis sentimen, atau pencarian informasi, dengan menghapus lapisan atasnya dan menggantinya dengan lapisan khusus sesuai dengan tugas yang ingin diselesaikan.

2.2.6 *Euclidean distance*

Euclidean distance adalah salah satu metrik paling dasar dan umum digunakan dalam matematika, ilmu komputer, dan statistik untuk mengukur jarak antara dua titik dalam ruang *Euclidean*. Jarak ini merepresentasikan panjang dari garis lurus yang menghubungkan dua titik tersebut. Euclidean distance sangat sederhana dan bergantung pada akar kuadrat dari jumlah perbedaan antara koordinat titik-titik di setiap dimensi.

Dalam ruang dua dimensi, jarak *Euclidean* antara dua titik $P(x_2 - x_1)^2$ dan $Q(x_2 - y_2)$ diberikan oleh rumus:

$$d(P,Q) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (6)$$

Untuk ruang tiga dimensi, rumusnya adalah:

$$d(P,Q) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \quad (7)$$

Secara umum, dalam ruang n-dimensi, jarak Euclidean antara dua titik $P(x_1, x_2, \ldots, x_n)$ dan $Q(y_1, y_2, \ldots, y_n)$ dihitung sebagai :

$$d(P,Q) = \sqrt{\sum_{i=1}^{n}(y_1 - x_i)^2} \quad (8)$$

Disini i adalah indeks yang mewakili setiap dimensi dalam ruang n-dimensi.

2.3 Metode Penyelesaian Masalah

2.3.1 State of The Art penelitian

Tabel 7. Matriks Jurnal Penelitian Terkait

No	Judul Karya Ilmiah, Nama, Tahun Terbit dan Penerbit	Objek dan Permasalahan	Metode Penyelesaian	Kinerja
1	**Judul :** Analisis kesesuaian jurnal ilmiah berbasis *Natural Language Processing* **Penulis :** Sitti mawaddah umar **Tahun :** 2023	**Objek :** Dokumen Jurnal ilmiah **Permasalahan:** analisis konsistensi penulisan jurnal ilmiah berdasarkan konten jurnal	*Natural Language Processing* (NLP), *Term Frequency – Inverse Document Frequency* (TF-IDF), *Cosine Similarity, Bidirectional Encoder Representations from Transformers* (BERT).	Pengujian sistem menunjukkan nilai kesamaan atau *coherence score* 0,740402.
2	**Judul :** *A correlation-based for classifying technical articles* **Penulis:** Kilany R, Ammar R, Rajasekaran S **Tahun :** 2020 **Penerbit :** IEEE	**Objek :** Artikel **Permasalahan :** alat komputasi yang mampu mengklasifikasi artikel menjadi 2 jenis yaitu paper-1 yang memiliki informasi yang menarik dan paper ke-2 tidak.	Menggunakan algoritma *text mining*	Hasil klasifikasi menggunakan metode *text mining* yang mengahasilkan korelasi paper positif 63% dan yang *negative* 58,3%
3	**Judul :** *A survey of topic models in text classification* **Penulis:** Xia L, Luo D, Zhang C, et al. **Tahun :** 2019 **Penerbit :** ICAIBD	**Objek :** Jurnal ilmiah **Permasalahan:** Bagaimana analisis proses pembuatan dokumen dan mengilustrasikan model grafis untuk setiap model topik	Menggunakan *Bag of Words* (BOW), analisis Laten (LSA), dan SVM.	Penelitian yang dilaksanakan mengambil data dengan mengklasifikasi dokumen dengan topik yang digambarkan dalam bentuk plot grafik menghasilkan

No	Judul Karya Ilmiah, Nama, Tahun Terbit dan Penerbit	Objek dan Permasalahan	Metode Penyelesaian	Kinerja
				akurasi sebesar 54%.
4	**Judul :** *Text Summarization Generation Based on Semantic Similarity* **Penulis:** Jinjing Chen dan Fucheng You **Tahun :** 2020 **Penerbit :** ICITBS	**Objek :** Abstrak dokumen **Permasalahan :** Apakah model sequence to sequence mampu menghasilkan rangkuman dan kesamaan makna dalam sebuah abstrak yang terdapat dalam dokumen.	*Natural Language Processing* (NLP), *sequence to sequence model, semantic similarity.*	Ukuran kesamaan untuk menghitung dapat meminimalkan kerugian fungsi dan menghasilkan kesamaan yang lebih baik.
5	**Judul :** *Journal Name Extraction from Japanese Scientific News Articles* **Penulis:** Kikuchi MYoshida Mumemura K. **Tahun :** 2018 **Penerbit :** APSIPA ASC	**Objek:** Artikel berita ilmiah di Jepang **Permasalahan:** Bagaimana mengidentifikasi nama jurnal dengan menggunakan kontek kejadian dalam artikel berita.	Menggunakan *generating training data, extracting journal names,* dan *evaluating the results.*	dalam penelitian ini menghasilkan kinerja yang secara umum baik dalam mengekstrak nama jurnal dari artikel berita ilmiah Jepang. Kinerja dievaluasi menggunakan *presisi*, *recall*, dan *F-measure*, dan hasilnya sebesar 84%.
6	**Judul :** *Editor Matching For Academic Journals Through Rich Semantic Network Development* **Penulis:** Prominski J, Shah P, Alvarado R **Tahun :** 2018 **Penerbit :** IEEE	**Objek :** Jurnal akademik **Permasalahan :** Bagaimana pengukuran antara editor dan dokumen dimodelkan sebagai node dalam grafik	Metode penyelasaian yang digunakan ialah dengan menggunakan algoritma NLP, metode text mining dan tematik korpus	Kinerja penelitian ini menghasilkan sistem rekomendasi pencarian kata dengan optimasi waktu 1,68/detik berdasarkan kutipan menggunakan faktor sitasi dan *co-reference*.
7	**Judul :** *Identifyng trends in*	**Objek :** Tren Artikel ilmu	Menggunakan metode TF-IDF	68,74% dapat digunakan

No	Judul Karya Ilmiah, Nama, Tahun Terbit dan Penerbit	Objek dan Permasalahan	Metode Penyelesaian	Kinerja
	data science articles using text mining **Penulis:** Adil S, Ebrahim M, Ali S,et al **Tahun :** 2019 **Penerbit :** IEEE	data **Permasalahan :** Bagaimana menemukan kata kunci dari dokumen	dan *Latent Dirichlet Allocation* (LDA)	sebagai pencarian kata kunci (*Key word)* dalam makalah.
8	**Judul :** *Keyword extraction and clustering for document* **Penulis:** Habibi M, Popescu-Belis A **Tahun :** 2019 **Penerbit :** IEEE	**Objek :** Rekomendasi dokumen **Permasalahan**: Bagaimana mengekstraksi dokumen dengan menggunakan kata kunci yang berpotensi terkait beberapa topik.	Menggunakan metode korpus dan TF-IDF	Akurasi yang dihasilkan metode TF-IDF dengan *output scoring* yang menghasilkan 69% kualitas baik dari kata kunci
9	**Judul :** *Modified TF-IDF term weighting strategies for text categorization* **Penulis:** Roul R, Sahoo J, Arora K **Tahun :** 2020 **Penerbit :** IEEE	**Objek :** Teks dokumen **Permasalahan:** Apa yang menjadi strategi dalam pemrosesan teks dalam dokumen	Menggunakan metode NLP, TF-IDF dan *vektor space* model	Dari 17,582 kata metode *vector space* model menghasilan 13,531 kosa kata dari dokumen.
10	**Judul :** *Implementation of fuzzy C-Means algorithm and TF-IDF on English Journal Summary* **Penulis:** Irfan M, Jumadi, Zulfikar W, et al **Tahun :** 2020 **Penerbit :** ICIC	**Objek :** Jurnal Bahasa inggris **Permasalahan :** Seberapa penting kalimat dapat diringkas dalam sebuah dokumen yang diberi pembobotan untuk penerapan metode Fuzzy K-means dan TF-IDF	Menggunakan metode TF-IDF dan Fuzzy K-Means	Hasil ringkasan sistem mendapatkan 65% dari tahap evaluasi *recall* tertinggi dengan presisi 47% sedangkan nilai terendah mendapatkan 27% dengan presisi 21%.
11	**Judul :**	**Objek :**	Menggunakan	Algoritma yang

No	Judul Karya Ilmiah, Nama, Tahun Terbit dan Penerbit	Objek dan Permasalahan	Metode Penyelesaian	Kinerja
	Clustering articles in Bahasa Indonesia *using self-organizing map* **Penulis :** Gunawan D, Amalia A, Charisma I **Tahun :** 2020 **Penerbit :** IEEE	Artikel Bahasa Indonesia **Permasalahan:** Sulitnya mengelompokkan artikel berbahasa Indonesia karena jumlahnya yang sangat banyak sehingga memakan waktu yang cukup lama.	metode TF-IDF, *Self-Organizing Map (*SOM), dan *Multi Word Expression.*	diterapkan untuk implementasikan artikel yang menggunakan bahasa Indonesia diatur untuk iterasi 100 kali dengan tingkat pembelajaran 60%
12	**Judul:** *Recommendation feature of scientific articles on open journal system using content-based filtering* **Penulis :** Setiadi H, Saptono R, Anggrainingsih R, et al. **Tahun :** 2019 **Terbit :** IEEE	**Objek :** Artikel jurnal **Permasalahan:** Bagaimana menanamkan fitur rekomendasi untuk menyarankan artikel yang relevan bagi pengguna berdasarkan kesamaan isi data.	Menggunakan metode TF-IDF, *Cosine Similarity* dan Clustering K-Means	Hasil nilai presisi metode *clustering k-means* adalah 68% dan metode *Cosine Similarity* rata-rata *score* presisinya adalah 54,15%
13	**Judul :** *Summarization of odia text document using Cosine Similarity and clustering* **Penulis :** Pattnaik S, Nayak A **Tahun :** 2019 **Penerbit :** IEEE	**Objek :** Dokumen teks **Permasalahan :** Bagaimana cara meringkas teks otomatis berdasarkan sub bidang.	Menggunakan metode NLP dan *Cosine Similarity*	Sistem yang dibangun untuk evaluasi bahasa Odia didapatkan *score* ringkasan kata 60,2%
14	**Judul :** *A Method for Thematic and Structural Visualization of Academic Content* **Penulis :** Amigud AArnedo-Moreno JDaradoumis	**Objek :** Tesis dan artikel ilmiah **Permasalahan :** bagaimana membuat pengamatan konten tertulis lebih efisien	Menggunakan metode *themetrack, corpus, ngram, text mining.*	Kinerja yang dihasilkan mencoba sampel sebanyak 20 tesis mahasiswa dan 20 artikel jurnal yang diterbitkan. Teks-teks tersebut

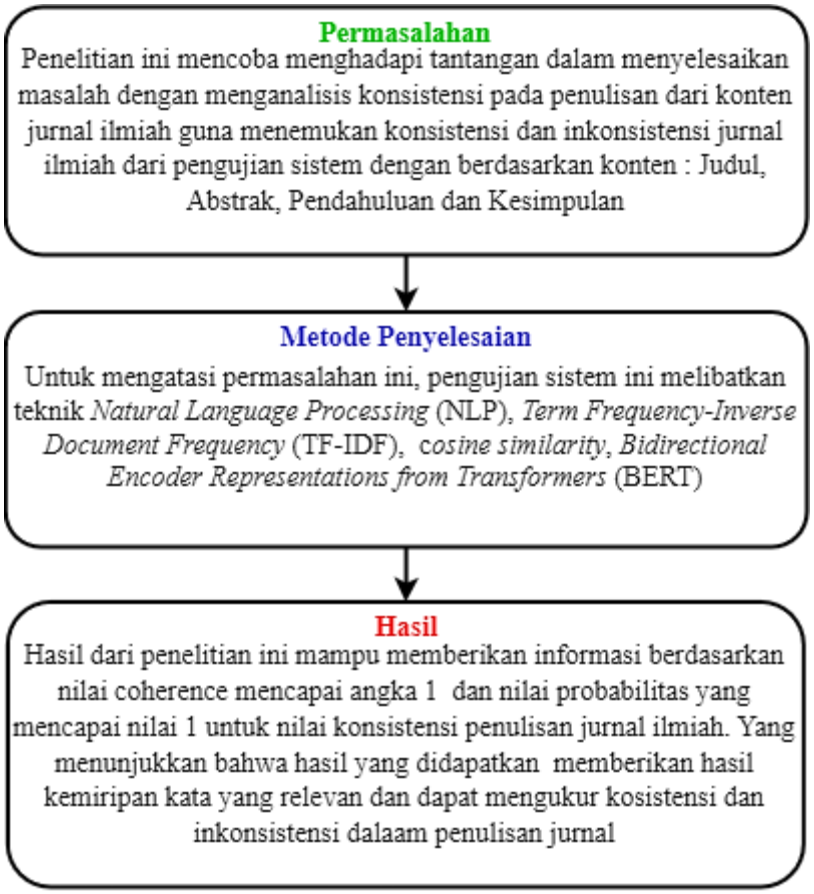
No	Judul Karya Ilmiah, Nama, Tahun Terbit dan Penerbit	Objek dan Permasalahan	Metode Penyelesaian	Kinerja
	T et al **Tahun :** 2017 **Penerbit :** IEEE			berkisar antara 3,000 hingga 85,000 kata, hasil akurasi yang didapatkan sebesar 82,73%
15	**Judul :** *Statutes Recommendation based on text similarity* **Penulis:** Zeng J, Ge J, Zhou Y, et al. **Tahun :** 2017 **Penerbit :** IEEE	**Objek :** Dokumen artikel **Permasalahan:** Bagaimana mendapatkan dokumen menggunakan kesamaan kata dengan cara statistik dan mengabaikan potensial semantic.	Metode TF-IDF dan (*Laten Analisis Semantic)* LSA	Bersama mencari kesamaan terhadap kata dengan penentuan frekuensi yang sama berdasarkan pembobotan yang dilakukan algoritma TF-IDF.

2.3.2 Metode Yang Diusulkan

Berdasarkan tabel *state of the art* metode yang diusulkan penelitian ini mampu untuk menganalisis konsistensi penulisan konten jurnal ilmiah dengan melibatkan beberapa teknik *Natural Language Processing* (NLP) dan *machine learning* (ML) untuk memastikan alur pemikiran dan informasi yang disajikan konsisten dan mudah dipahami. Pertama, dilakukan preprocessing teks untuk mempersiapkan teks agar dapat dianalisis lebih lanjut. Selanjutnya, metode *Term Frequency-Inverse Document Frequency* (TF-IDF) digunakan untuk mengubah teks menjadi vektor numerik yang mencerminkan pentingnya kata dalam dokumen relatif terhadap kumpulan dokumen lainnya. Algoritma BERT juga ditambahkan dalam penelitian ini sebagai metode yang mampu menangkap kontek teks lebih dalam, sebab *Cosine Similarity* hanya menangkap kesamaan kata namun apabila ada kata yang memiliki makna yang sama tapi dengan kata yang berbeda tidak dapat ditemukan. Jadi, *Cosine Similarity* berfungsi mengukur kesamaan dan konsistensi antar bagian jurnal seperti judul, abstrak, pendahuluan, dan kesimpulan, digunakan *Cosine Similarity*. Metode ini diharapkan dapat

meningkatkan kualitas penulisan jurnal ilmiah dengan memastikan konsistensi dan meningkatkan pemahaman pembaca terhadap isi jurnal.

2.4 Kerangka Pikir Penelitian



Gambar 4. Kerangka Pikir Penelitian