AICOMS
Applied Information Technology and Computer Science
e-ISSN: 2964-7703
Vol. 4 No. 2, Desember-2025, pp. 01-09

Keyword Extraction Abstrak Jurnal Ilmiah Menggunakan Metode TF-IDF dan KeyBERT

Rayvin Suhartoyo [1], Valen Julyo Armando Davincy Lin [2], Hafiz Irsyad [3], and Abdul Rahman [4]

[1] Universitas Multi Data Palembang; rayvinsuhartoyo_2226250007@mhs.mdp.ac.id
[2] Universitas Multi Data Palembang; valenjulyoarmandodavincylin_2226250045@mhs.mdp.ac.id
[3] Universitas Multi Data Palembang; hafizirsyad@mhs.mdp.ac.id
[4] Universitas Multi Data Palembang; arahman@mhs.mdp.ac.id
* Universitas Multi Data Palembang; rayvinsuhartoyo_2226250007@mhs.mdp.ac.id

Info Artikel:
Dikirim: 30 Mei 2025
Direvisi: 02 Juli 2025
Diterima: 02 Juli 2025

Abstract: Keyword extraction is a significant technique in natural language processing (NLP) that serves to summarize the essence of a document, such as a scientific journal summary. This study aims to analyze the effectiveness of two keyword extraction methods, namely Term Frequency-Inverse Document Frequency (TF-IDF) and KeyBERT, in finding significant keywords from a collection of scientific journal abstracts. The dataset used consists of several scientific journal abstracts accompanied by manual keywords as a basis for assessment. The TF-IDF method relies on the frequency of words in the document, while KeyBERT utilizes a cosine similarity approach based on the BERT transformer model to determine the most meaningful keywords. The research findings show that the KeyBERT method and the TF-IDF method have a moderate level of similarity with semantic similarity values of 0.578 for the KeyBERT method and 0.469 for the TF-IDF method, respectively. These results show significant potential for the use of machine learning and deep learning-based models with both methods for topic classification systems, especially in the fields of information retrieval and text mining.

Keywords: Cosine Similarity; Information Retrieval; Keyword Extraction; KeyBERT; Semantic Similarity; TF-IDF.

Intisari: Ekstraksi kata kunci adalah teknik signifikan dalam pemrosesan bahasa alami (NLP) yang berfungsi untuk merangkum esensi dari sebuah dokumen, seperti ringkasan jurnal ilmiah. Penelitian ini bertujuan untuk menganalisis efektivitas dua metode ekstraksi kata kunci, yakni Term Frequency-Inverse Document Frequency (TF-IDF) dan KeyBERT, dalam menemukan kata kunci signifikan dari koleksi abstrak jurnal ilmiah. Kumpulan data yang dipakai terdiri dari beberapa abstrak jurnal ilmiah yang disertai dengan kata kunci manual sebagai dasar penilaian. Metode TF-IDF bergantung pada frekuensi kata dalam dokumen, sementara KeyBERT memanfaatkan pendekatan *cosine similarity* yang didasarkan pada model transformer BERT untuk menentukan kata kunci yang paling tepat secara makna. Temuan penelitian menunjukkan bahwa metode KeyBERT dan metode TF-IDF memiliki tingkat kemiripan sedang dengan nilai semantic similarity masing-masing adalah 0,578 untuk metode KeyBERT dan 0,469 untuk metode TF-IDF. Hasil ini menunjukkan potensi signifikan penggunaan model berbasis *machine learning* dan *deep learning* dengan kedua metode untuk sistem klasifikasi topik, terutama dalam bidang *information retrieval* dan *text mining*

Kata Kunci: *Cosine Similarity*; *Information Retrieval*; *Keyword Extraction*; KeyBERT; *Semantic Similarity*; TF-IDF.

1. Pendahuluan

Karya Ilmiah merupakan salah satu bentuk sumbangsih yang nyata untuk dunia pendidikan. Ekosistem pendidikan dapat terus mengalami kemajuan karena aktivitas - aktivitas penelitian masih terus berjalan. Penelitian ini kemudian dibuktikan dan didokumentasikan dalam bentuk karya ilmiah untuk menjadi referensi bagi peneliti lain. Salah satu bentuk karya ilmiah yang populer di lingkungan Perguruan Tinggi adalah Artikel ilmiah baik yang akan di publikasi pada Jurnal Ilmiah maupun prosiding ilmiah [1].

Saat ini, identifikasi kata kunci untuk artikel ilmiah masih dilakukan dengan cara manual. Hal ini mengakibatkan proses pemilihan kata kunci menjadi kurang efisien dan memakan waktu, terutama jika jumlah artikel yang ada sangat besar. Tidak semua penulis dapat menciptakan kata kunci untuk tulisan mereka, dan tidak semua kata kunci dapat secara akurat mencerminkan konten teks karena adanya subjektivitas manusia, sehingga kata kunci yang dihasilkan kurang bersifat umum [2]. Salah satu alternatif yang memudahkan pengguna untuk mendapatkan informasi penting adalah dengan diadakan sistem pengekstrasian kata kunci yang diterapkan pada aplikasi *search engine*. Selain itu hasil dari pengekstraksian kata kunci juga dapat dijadikan rekomendasi kata kunci pencarian yang dibutuhkan untuk mendukung kata kunci utama [3].

Metode *Term Frequency-Inverse Document Frequency* (TF-IDF) adalah teknik yang digunakan untuk menentukan bobot suatu istilah terkait dengan sebuah dokumen [4]. Walaupun TF-IDF bisa mengidentifikasi kata-kata penting berdasar frekuensi hadirnya, metode ini sering kali kurang menangkap makna semantik dan konteks dari kata-kata tersebut. Berdasarkan penelitian yang dilakukan oleh Mihaelcea dan Tarau (2004) memanfaatkan algoritma TF-IDF sebagai komponen dalam algoritma TextRank untuk merangkum dan mengekstrak informasi penting dari sebuah teks [5]. Sementara itu penelitian yang dilakukan Hulth (2003) mengkaji mengenai kinerja TF-IDF untuk ekstraksi kata kunci dan mendapatkan hasil yang cukup baik meskipun tidak mengandung pemahaman secara semantik [6]. Penelitian lain oleh lahiri (2014) juga menggunakan metode TF-IDF untuk ekstraksi topik dari artikel ilmiah menekankan pentingnya pemrosesan bahasa alami dalam tahap preprocessing [7].

Untuk mengatasi kelemahan TF-IDF, berbagai metode berbasis representasi semantik telah diciptakan, salah satunya adalah KeyBERT. KeyBERT merupakan model pengambilan kata kunci yang menggunakan *Bidirectional Encoder Representations from Transformers* (BERT) dan dapat memahami konteks kalimat, serta menghasilkan ekstraksi kata kunci yang lebih relevan dan bermakna secara semantik. Penelitian yang dilakukan oleh Grootendorst (2020), sebagai pengembang utama KeyBERT, membuktikan metode ini secara konsisten menghasilkan kata kunci yang lebih relevan dan kontekstual dibandingkan metode statistik tradisional [8]. Penelitian yang dilakukan Reimers dan Gurevych (2019) melalui pengembangan Sentence-BERT turut memperkuat fondasi KeyBERT dalam efisiensi representasi antar kalimat [9].

Berdasarkan penelitian terdahulu, masih banyak penelitian ekstraksi kata kunci yang masih menggunakan evaluasi berbasis kemiripan string atau token literal (*accuracy*, *precision*, *recall*, and *F1 score*), tanpa mengukur kesesuaian semantik antara hasil ekstraksi dan konteks dokumen. Hal ini menyisakan ruang untuk pendekatan evaluasi berbasis *semantic similarity*, yang dapat memberikan gambaran objektif mengenai relevansi makna dari kata kunci yang dihasilkan terhadap konten dokumen. Selain itu, sebagian besar studi sebelumnya menggunakan dataset umum seperti berita, wikipedia, atau blog. Masih sedikit yang melakukan ekstraksi kata kunci pada abstrak jurnal ilmiah, khususnya dalam domain *Information Retrieval*. Sebagian pendekatan juga hanya melakukan *text preprocessing* saja dalam mengolah dokumennya. Padahal, penggunaan algoritma seperti TextRank sebagai metode peringkasan otomatis dapat membantu menyaring kalimat penting, sehingga kata kunci yang diekstrak lebih fokus dan relevan dengan isi dokumen.

Penelitian ini bertujuan untuk melihat performa metode TF-IDF dan KeyBERT dalam mengekstraksi kata kunci dari abstrak jurnal ilmiah. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan sistem pencarian informasi ilmiah yang lebih efisien dan akurat.

2. Tinjauan Pustaka

Bagian ini menyajikan landasan teori yang mendukung metode yang diajukan untuk menyelesaikan suatu permasalahan serta pengembangan metode tersebut, yang didasarkan pada referensi yang kredibel seperti buku, jurnal, prosiding, dan artikel ilmiah lainnya.

2.1 *TF-IDF(Term Frequency – Inverse Document Frequency)*

TF-IDF adalah teknik statistik yang digunakan dalam pemrosesan bahasa alami dan pencarian informasi untuk menilai pentingnya suatu kata dalam sebuah dokumen relatif terhadap kumpulan dokumen lainnya. Metode ini menggabungkan dua komponen utama: frekuensi kemunculan kata dalam dokumen (*Term Frequency* atau TF) dan kebalikan dari frekuensi dokumen yang mengandung kata tersebut (*Inverse Document Frequency* atau IDF). TF mengukur seberapa sering sebuah kata muncul dalam dokumen tertentu, sementara IDF menilai

seberapa umum atau langka kata tersebut di seluruh korpus. Dengan mengalikan TF dan IDF, TF-IDF memberikan bobot yang lebih tinggi pada kata-kata yang sering muncul dalam dokumen tertentu tetapi jarang muncul di dokumen lain, sehingga membantu dalam menyoroti kata-kata yang lebih informatif dan relevan dalam konteks dokumen tersebut [4].

2.2 *KeyBERT*

KeyBERT adalah sebuah metode ekstraksi kata kunci yang memanfaatkan kekuatan model bahasa berbasis transformer, khususnya BERT (*Bidirectional Encoder Representations from Transformers*). KeyBERT menggunakan *embedding* dari BERT untuk merepresentasikan dokumen dan kandidat kata kunci dalam bentuk vektor numerik yang kaya konteks semantik. Dengan cara ini, KeyBERT dapat menangkap hubungan makna antara kata-kata dalam teks, tidak hanya sekedar frekuensi kemunculan kata seperti pada metode tradisional TF-IDF [10]. Untuk mengembangkan ekstraktor kata kunci yang andal, kami menggunakan KeyBERT, yang memanfaatkan model BERT untuk menghasilkan representasi vektor dari dokumen dan kandidat frasa, lalu membandingkannya menggunakan kesamaan kosinus. Frasa dengan skor tertinggi dipilih sebagai kata kunci utama. Meski efektif untuk ekstraksi umum, akurasi KeyBERT menurun di domain khusus karena model defaultnya mengandalkan konteks bahasa Inggris umum. Misalnya, frasa seperti "hipotesis ilmiah" mungkin dianggap penting dalam abstrak fisika, padahal sebenarnya umum dan kurang informatif dalam konteks tersebut. Oleh karena itu, penyesuaian terhadap karakteristik domain sangat diperlukan [11].

2.3 *Dataset*

Dataset merupakan komponen penting dalam proses ekstraksi kata kunci karena berfungsi sebagai sumber data utama yang digunakan untuk pelatihan, pengujian, maupun evaluasi algoritma seperti TF-IDF dan KeyBERT. Dalam konteks abstrak jurnal ilmiah, dataset biasanya terdiri dari kumpulan teks abstrak yang dilengkapi dengan kata kunci manual yang ditentukan oleh penulis atau editor jurnal. Kualitas dan karakteristik dataset sangat mempengaruhi hasil ekstraksi kata kunci. Beberapa faktor penting meliputi kualitas anotasi kata kunci, jumlah dan keberagaman data, bahasa yang digunakan, serta struktur teks. Dataset yang memiliki anotasi kata kunci yang konsisten dan relevan akan memberikan dasar evaluasi yang baik bagi metode ekstraksi otomatis. Selain itu, penggunaan dataset yang luas dan berasal dari berbagai bidang ilmu dapat meningkatkan kemampuan generalisasi model. Bahasa dalam dataset juga perlu diperhatikan karena algoritma seperti KeyBERT bergantung pada model bahasa tertentu, misalnya BERT berbahasa Inggris atau IndoBERT untuk bahasa Indonesia. Oleh karena itu, pemilihan dataset yang tepat dan relevan sangat menentukan keberhasilan dalam menghasilkan kata kunci yang akurat dan bermakna dari sebuah abstrak jurnal ilmiah.

2.4 *TextRank*

TextRank merupakan algoritma yang digunakan untuk memperoleh kata-kata paling signifikan dalam suatu dokumen teks. TextRank yang didasarkan pada graf digunakan untuk memberikan peringkat pada teks, dan kalimat-kalimat dalam teks direpresentasikan sebagai simpul atau titik dalam grafik. Peneliti lain juga menyatakan bahwa metode TextRank adalah algoritma ringkasan yang berbasis graf, yang dirancang berdasarkan metode PageRank, terdiri dari vertex yang mewakili kalimat dalam dokumen dan edge yang menunjukkan hubungan kesamaan antar kalimat [12].

2.5 *Semantic Similarity*

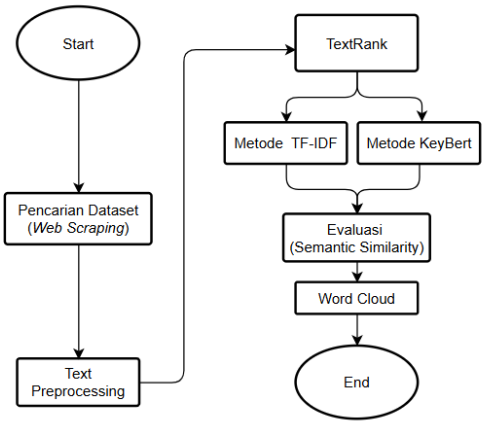
Semantic Similarity merupakan sebuah metrik yang didefinisikan dalam sekumpulan dokumen, di mana gagasan mengenai jarak antar dokumen tersebut mengacu pada kesamaan makna atau konten semantik yang berbeda dengan kesamaan yang dapat dihitung berdasarkan representasi sintaktik (atau format string dari dokumen-dokumen itu). Metode ini adalah alat matematis yang digunakan untuk memperkirakan kekuatan dari relasi semantik antara unit bahasa, konsep, melalui deskripsi numerik yang diperoleh berdasarkan perbandingan antara informasi yang mendukung maknanya atau mendeskripsikan kata pokoknya. Kemiripan semantik biasanya disamakan dengan keterkaitan semantik. Keterkaitan semantik mencakup hubungan antara dua istilah, sedangkan kemiripan semantik hanya mencakup hubungan “adalah”. Misalnya, “mobil” memiliki kesamaan dengan “angkot”, namun berkaitan dengan “jalan” dan “mengemudi” [13].

3. Metode dan Hasil Penelitian

3.1 *Metode Penelitian*

Dalam penelitian ini, terdapat rancangan tahapan dimulai dari proses *web scraping* untuk mendapatkan dataset yang asli. Data yang menjadi objek penelitian ini adalah abstrak jurnal ilmiah yang berhubungan dengan tema *information retrieval.* Selanjutnya akan dilakukan proses *text preprocessing* untuk menghasilkan dataset yang lebih terstruktur. Selanjutnya, dilakukan proses pencarian kalimat penting dengan metode TextRank untuk meringkas abstrak jurnal ilmiah menjadi kalimat penting yang kemudian akan dilakukan proses *keyword*

extraction dengan metode TF-IDF(*Term Frequency - Inverse Data Frequency*) dan metode KeyBERT untuk mencari kata kunci dari abstrak jurnal ilmiah tersebut. Hasil dari rancangan ini dapat dilihat pada Gambar 1.



Gambar 1. Flowchart keyword extraction

3.1.1 Data Preparation

Dalam penelitian ini, pengumpulan data dilakukan dengan cara *web scraping*. Web Scraping adalah sebuah cara pengambilan suatu data atau informasi tertentu dengan jumlah besar untuk nantinya digunakan dalam berbagai keperluan seperti riset, analisis, dan lainnya [14]. Penelitian ini mengambil data menggunakan API dari tiga website yaitu, Semantic Scholar, Arxiv, dan DOAJ. Fokus penelitian ini adalah abstrak jurnal ilmiah dengan 5 kata kunci pencarian yaitu, *information retrieval, text mining, document ranking, query expansion, tf-idf* dengan fokus jurnal ilmiah dari tahun 2000 - 2024. Kata kunci ini digunakan untuk mencari jurnal ilmiah yang berhubungan dengan *information retrieval* yang merupakan fokus penelitian. Hasil *web scraping* pada penelitian ini berjumlah sebanyak 826 jurnal ilmiah yang didapatkan dari ketiga *website* tersebut.

3.1.3 TextRank

TextRank merupakan *graph-based ranking algorithm* (graf dengan model peningkatan) untuk pemrosesan teks dari dokumen bahasa alami atau manusia. Dalam penelitian ini, Textrank yang digunakan adalah *TextRank for keyword extraction* (ekstraksi kata kunci)[15]. TextRank ini digunakan untuk menyeleksi abstrak jurnal ilmiah menjadi beberapa kalimat - kalimat penting yang memudahkan untuk melakukan ekstraksi kata kunci nantinya.

3.1.3 Text Preprocessing

Text Preprocessing adalah sebuah proses yang dilakukan untuk mengubah bentuk data teks yang belum terstruktur menjadi data teks yang terstruktur sesuai dengan kebutuhan, untuk proses analisis yang lebih lanjut[16]. Dalam penelitian ini, dilakukan beberapa tahapan *text preprocessing* untuk mengelola data abstrak jurnal ilmiah yang telah dikumpulkan melalui *web scraping*. Tahapan - tahapan *text preprocessing* adalah:

1. Case Folding
   Case Folding adalah teknik yang bertujuan untuk menyamakan penggunaan huruf besar atau kecil dalam teks dengan mengubah semua huruf menjadi huruf kecil(*Lowercase*) atau huruf besar (*Uppercase*). Seluruh abstrak jurnal ilmiah yang didapatkan akan diubah menjadi huruf kecil semuanya sebelum dikelola lebih lanjut.
2. Filtering / Stopword
   Filtering adalah teknik yang bertujuan untuk menghapus kata-kata yang tidak relevan dalam data teks, seperti kata sambung, kata depan, kata ganti dan lain sebagainya. *Filtering* dilakukan dalam data abstrak jurnal ilmiah untuk menghilangkan kata-kata yang tidak terpakai dalam sebuah kalimat dan *special character* yang ada dalam dataset. Setelah dataset melalui proses *tokenization*, kalimat - kalimat tersebut akan dipecah lagi menjadi kata-kata yang dianggap penting (*Stopword*).
3. Tokenization

Tokenization merupakan teknik yang bertujuan untuk memecah teks menjadi unit-unit yang lebih kecil, seperti kata, frasa atau kalimat. Dalam penelitian ini, abstrak akan dipecah menjadi beberapa kalimat yang kemudian akan diproses kembali menjadi kata - kata penting.

4. Lemmatization
   Lemmatization merupakan teknik yang bertujuan untuk mengubah kata-kata yang memiliki bentuk berbeda tapi memiliki makna yang sama menjadi kata baku. Teknik ini mirip dengan *Stemming*, hanya saja lebih halus dan menghasilkan kata baku yang lebih relevan dan sesuai. Dalam penelitian ini, teknik ini digunakan pada kata - kata penting sebelumnya untuk mempermudah kinerja dari proses *keyword extraction* nantinya.

3.1.4 Keyword Extraction dengan TF-IDF

TF-IDF (*Term Frequency - Inverse Document Frequency*) adalah teknik yang bertujuan untuk mengevaluasi seberapa penting suatu kata(term) dalam sebuah dokumen dalam konteks koleksi dokumen yang lebih besar [17],[18]. Nilai TF-IDF didapat dari perkalian antara nilai TF yang dikalikan dengan nilai IDF dan dapat dihitung dengan Persamaan (1).

$$TF - IDF = TF(D, DF) * IDF\left(\frac{D}{DF}\right) = \frac{D}{DF} * \log_{10}\left(\frac{D}{DF}\right) \quad (1)$$

Ket:
D = Jumlah dokumen
DF = Jumlah kata yang muncul pada sebuah dokumen

Dalam metode TF-IDF terdapat dua proses dalam pembobotan pada setiap kata.

1. TF (*Term Frequency*)
   TF atau *term frequency* adalah jumlah kata yang muncul dalam sebuah dokumen. Untuk menghitung jumlah kata yang muncul dapat dilihat pada Persamaan (2).

$$TF(D, DF) = \frac{D}{DF} \quad (2)$$

2. IDF (*Inverse Document Frequency*)
   IDF atau *Inverse Document Frequency* adalah nilai seberapa penting kata dalam sebuah dokumen. Untuk menghitung nilai IDF dapat dilihat pada Persamaan (3).

$$IDF\left(\frac{D}{DF}\right) = \log_{10}\left(\frac{D}{DF}\right) \quad (3)$$

3.1.5 Keyword Extraction dengan KeyBERT

KeyBERT adalah teknik ekstraksi kata kunci yang minimal dan mudah digunakan yang memanfaatkan *embedding* BERT untuk membuat kata kunci dan frasa kunci yang paling mirip dengan suatu dokumen[8]. Dalam penelitian ini, metode KeyBERT akan membagi teks menjadi beberapa kandidat frasa menggunakan n-gram (1-3 kata) yang kemudian hasilnya akan ditransformasi (*embedding*) menjadi vektor. Dengan nilai vektor sebelumnya, akan dihitung nilai cosine similarity untuk melihat kemiripan dengan dokumennya yang kemudian kandidat kata kunci akan diurutkan berdasarkan nilai dari cosine similarity tertinggi masing masing. Persamaan untuk menghitung nilai cosine similarity dapat dilihat pada Persamaan (4).

$$Score(\vec{k}) = \cos(\vec{k}, \vec{d}) = \frac{\vec{k} \cdot \vec{d}}{\left|\left|\vec{k}\right|\right| \cdot \left|\left|\vec{d}\right|\right|} \quad (4)$$

Ket:
$\vec{k}$= Kandidat kata kunci
$\vec{d}$= Dokumen

3.1.6 Evaluasi

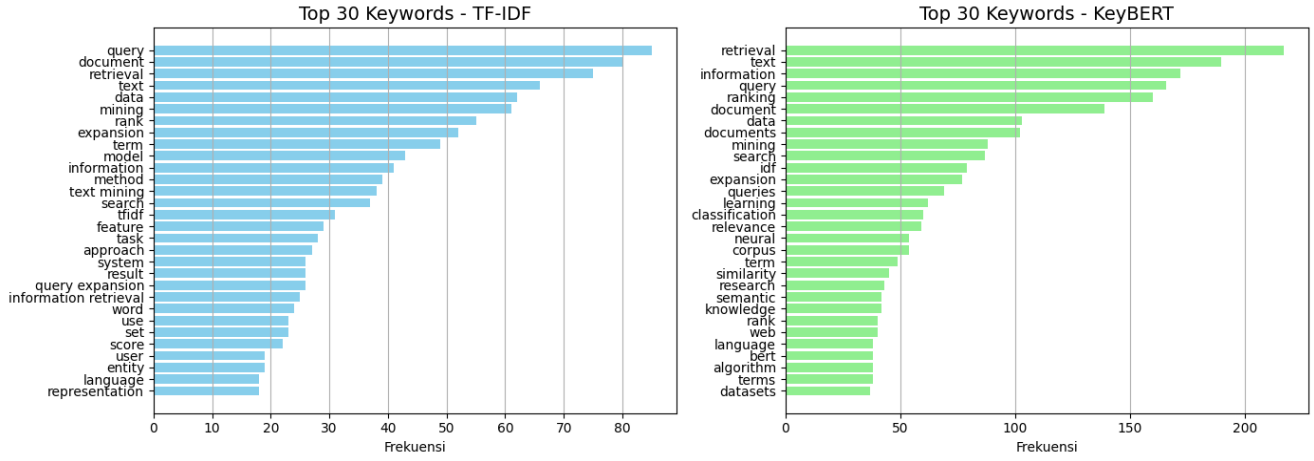
Teknik yang digunakan untuk melakukan evaluasi sistem pada penelitian ini menggunakan teknik Semantic Similarity. Semantic Similarity mengacu pada tingkat kesamaan antara kata-kata. Tujuan dari semantic similarity ini adalah untuk mengukur seberapa dekat hubungan atau analogi konsep, ide, atau informasi yang disampaikan dalam dua teks[19]. Dalam penelitian ini, semantic similarity bertujuan untuk membandingkan kata kunci hasil metode TF-IDF dan keyBERT untuk dibandingkan dengan kata kunci dari query pencarian yang digunakan saat dilakukannya web scraping untuk mengambil dataset. Interpretasi nilai dari semantic similarity dapat dilihat dari Tabel 1.

Tabel 1. Interpretasi Nilai *Semantic Similarity*

No	Interpretasi Nilai	Keterangan
1.	1,00 - 0,80	Sangat Mirip
2.	0,79 - 0,60	Cukup Mirip
3.	0,59 - 0,40	Sedang
4.	< 0,40	Tidak relevan

3.2 *Hasil Penelitian*

Dari 826 jurnal ilmiah yang didapatkan dari proses web scraping dari ketiga website yang digabungkan menjadi satu dataset dan kemudian akan dilakukan keyword extraction dengan metode TF-IDF dan metode KeyBERT. Didapatkan hasil top 30 kata kunci yang paling sering dibahas dalam jurnal ilmiah dengan tema *information retrieval* dengan fokus jurnal pada tahun 2000 - 2024 dapat dilihat pada Gambar 2.

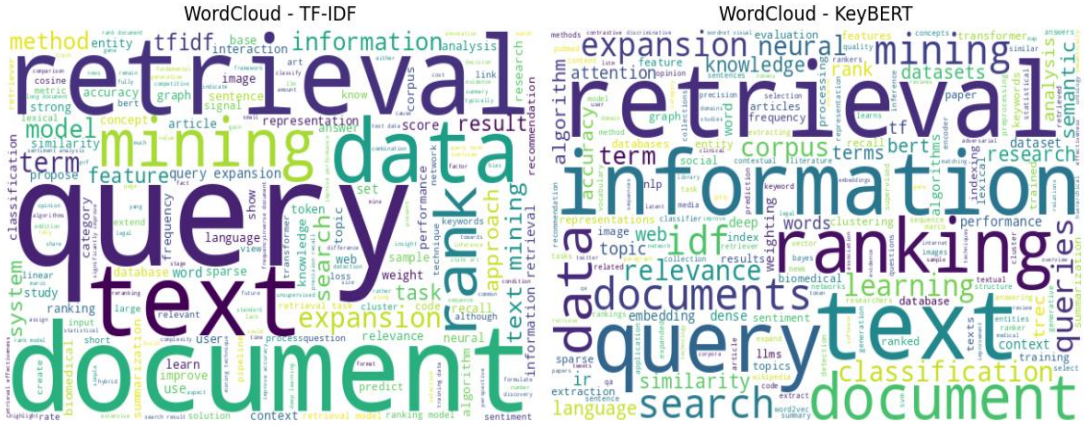


(a) (b)

Gambar 2. Visualisasi Keyword Extraction

(a) Visualisasi Keyword Extraction dengan metode TF-IDF; (b) Visualisasi Keyword Extraction dengan metode KeyBERT

Gambar 3 menunjukkan visualisasi keyword extraction dari kedua metode dengan memanfaatkan library word cloud. Gambar 3 merepresentasikan seluruh kata kunci yang didapatkan dari hasil *keyword extraction*. Semakin besar nilai frekuensi dari kemunculan kata kunci dalam abstrak jurnal ilmiah, maka akan semakin besar visualisasi *word cloud*-nya.



(a) (b)

Gambar 3. Visualisasi dengan library Word Cloud
(a) Visualisasi Library Word Cloud dengan metode TF-IDF; (b) Visualisasi Library Word Cloud dengan metode KeyBERT

Untuk melihat jumlah frekuensi dari top 10 kata kunci tertinggi hasil keyword extraction dari kedua metode, dapat dilihat pada Tabel 2 untuk metode TF-IDF dan Tabel 3 untuk metode KeyBERT. Jika merujuk pada Tabel 2, kata kunci tertinggi hasil keyword extraction dengan metode TF-IDF adalah Query. Kata kunci Query muncul sebanyak 85 kali dan pada metode KeyBERT, kata kunci Query muncul sebanyak 166 kali dari 826 abstrak jurnal ilmiah yang digunakan sebagai dataset. Pada tabel 3, dapat dilihat jika kata kunci tertinggi hasil keyword extraction dengan metode KeyBERT adalah Retrieval. Kata kunci Retrieval muncul sebanyak 217 kali dan dengan metode TF-IDF, kata kunci Retrieval muncul sebanyak 75 kali dari 826 abstrak jurnal ilmiah yang digunakan sebagai dataset.

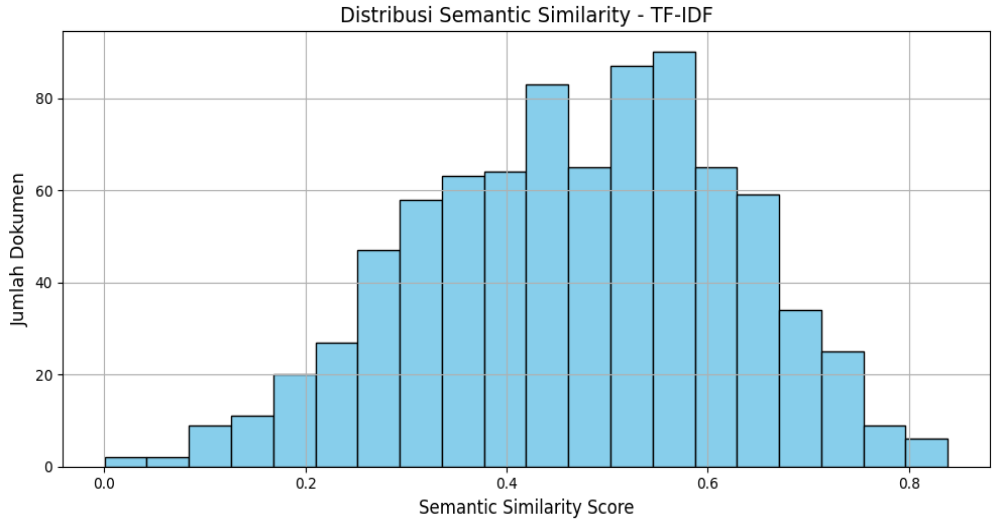
Tabel 2. Top 10 Nilai Frekuensi dari Keyword Extraction dengan metode TF-IDF

No	Kata Kunci	Frekuensi
1.	Query	85
2.	Document	80
3.	Retrieval	75
4.	Text	66
5.	Data	62
6.	Mining	61
7.	Rank	55
8.	Expansion	52
9.	Term	49
10.	Model	43

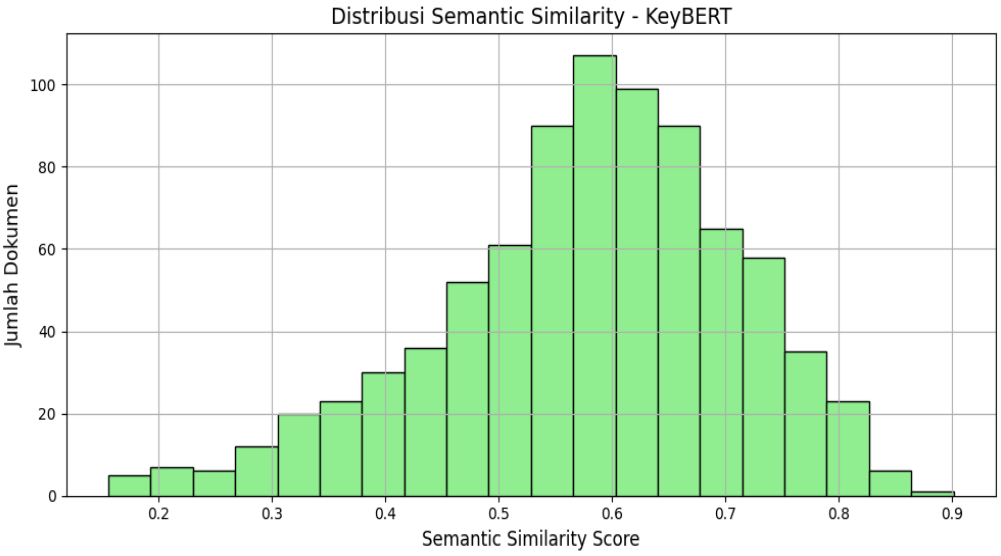
Tabel 3. Top 10 Nilai Frekuensi dari Keyword Extraction dengan metode KeyBERT

No	Kata Kunci	Frekuensi
1.	Retrieval	217
2.	Text	190
3.	Information	172
4.	Query	166
5.	Ranking	160
6.	Document	139
7.	Data	103
8.	Documents	102
9.	Mining	88
10.	Search	87

Untuk melihat seberapa besar tingkat keberhasilan metode TF-IDF dan metode KeyBERT dalam melakukan keyword extraction dari abstrak jurnal ilmiah. Dilakukan evaluasi dengan menghitung nilai rata - rata dari semantic similarity dari kedua metode yang digunakan. Nilai rata - rata dari semantic similarity dengan menggunakan metode TF-IDF adalah 0,469 dan nilai rata-rata dari semantic similarity dengan menggunakan metode KeyBERT adalah 0,578. Berdasarkan interpretasi dari nilai rata – rata semantic similarity pada Tabel 1. Kedua metode berada dalam rentang 0,59 – 0,4 yang dimana menunjukkan tingkat kemiripan sedang. Hal ini berarti bahwa kedua metode masih dapat dikembangkan kembali untuk meningkatkan tingkat kemiripan antara abstrak jurnal ilmiah dan query pencarian yang digunakan.



Gambar 4. Visualisasi Distribusi Semantic Similarity dengan metode TF-IDF



Gambar 5. Visualisasi Distribusi Semantic Similarity dengan metode KeyBERT

Dilihat dari visualisasi distribusi nilai semantic similarity pada Gambar 4 dan Gambar 5. Gambar 4 menunjukkan distribusi nilai dari semantic similarity menggunakan metode TF-IDF. Dari Gambar 4, dapat dilihat jika distribusi nilai semantic similarity berada dalam rentang 0,4 hingga 0,6. Hal ini menunjukkan banyak dokumen yang diekstrak menggunakan metode TF-IDF memiliki tingkat kemiripan sedang dengan query pencariannya. Metode TF-IDF berfokus pada jumlah frekuensi kemunculan kata dalam dokumen tersebut tanpa mempertimbangkan hubungan antar kata. Untuk grafik pada Gambar 5 mempresetasikan distribusi nilai semantic similarity dengan metode KeyBERT. Dari Gambar 5, dapat dilihat jika distribusi nilai semantic similarity berada dalam rentang 0,5 hingga 0,7. Hal ini menunjukkan banyak dokumen yang diektrak menggunakan metode TF-IDF memiliki tingkat kemiripan berada dalam rentang sedang hingga cukup mirip. Metode KeyBERT memiliki tingkat kesesuaian makna yang lebih luas karena metode KeyBERT mampu mengekstraksi kata kunci

yang tidak hanya relevan dari jumlah kata yang muncul tetapi juga relevan secara makna dan hubungan antar kata.

4. Kesimpulan

Hasil penelitian menunjukan bahwa tingkat kemiripan sedang dengan menggunakan pendekatan metode TF-IDF dan KeyBERT terhadap dataset abstrak jurnal yang didapatkan dengan nilai rata - rata *semantic similarity* untuk metode TF-IDF adalah 0,469 dan nilai rata-rata *semantic similarity* adalah 0,578. Dari hasil evaluasi tersebut, dapat disimpulkan bahwa metode KeyBERT dan TF-IDF cukup baik dalam melakukan *keyword extraction* namun masih bisa dioptimasi dengan penambahan metode lainnya. Selain itu, gabungan dari beberapa tahapan proses seperti *text preprocessing* dan TextRank terbukti efektif untuk diterapkan pada dataset sebelum memasuki proses *keyword extraction*.

Dengan adanya penelitian ini, diharapkan membuka potensi dan peluang untuk dilakukannya integrasi lanjutan dengan sistem *Machine Learning* dan *Deep Learning* untuk menguatkan analisis semantik dan interpretabilitas yang lebih luas lagi. Penelitian ini juga diharapkan dapat menjadi solusi dalam sistem klasifikasi topik, *indexing* jurnal, serta sistem rekomendasi yang berbasis konten ilmiah terutama dalam bidang *Information Retrieval* dan *Text Mining*.

Daftar Pustaka

[1] B. Bahar, "Pengembangan Model Sistem Informasi Manajemen Pengelolaan Artikel Ilmiah Berbasis Web Menggunakan Metode Extreme Programming," *Jutisi J. Ilm. Tek. Inform. dan Sist. Inf.*, vol. 9, no. 3, p. 1, 2021, doi: 10.35889/jutisi.v9i3.537.

[2] M. A. Shiddiq, "Ekstraksi Kata Kunci pada Artikel Menggunakan Metode Textrank," vol. 1, no. 1, pp. 1–97, 2019.

[3] A. Kurniawan, "Aplikasi sistem ekstraksi kata kunci berbahasa indonesia menggunakan algoritma textrank studi kasus data wikipedia Indonesia," *Repository.Uinjkt.Ac.Id*, 2021.

[4] R. Al Rasyid and D. H. U. Ningsih, "Penerapan Algoritma TF-IDF dan Cosine Similarity untuk Query Pencarian Pada Dataset Destinasi Wisata," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 8, no. 1, pp. 170–178, 2024, doi: 10.35870/jtik.v8i1.1416.

[5] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," *Proc. 2004 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2004 - A Meet. SIGDAT, a Spec. Interes. Gr. ACL held conjunction with ACL 2004*, vol. 85, pp. 404–411, 2004.

[6] M. Ciaramita and M. Johnson, "Supersense Tagging of Unknown Nouns in WordNet," *Proc. 2003 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2003*, pp. 168–175, 2003, doi: 10.3115/1119355.1119377.

[7] C. Argueta and Y. S. Chen, "Multi-Lingual Sentiment Analysis of Social Data Based on Emotion-Bearing Patterns," *Soc. 2014 - 2nd Work. Nat. Lang. Process. Soc. Media, conjunction with COLING 2014*, no. 101, pp. 38–43, 2014, doi: 10.3115/v1/w14-5906.

[8] M. Grootendorst, "KeyBERT: Minimal keyword extraction with BERT.," Zenodo. [Online]. Available: https://github.com/MaartenGr/KeyBERT

[9] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," *EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 3982–3992, 2019, doi: 10.18653/v1/d19-1410.

[10] A. Bert, "Applied Information Technology and Computer Science Sentimen Analisis Overclaim Skincare Skintific Menggunakan," vol. 3, no. 2, pp. 1–5, 2024.

[11] J. Newburn, "Keyword Extraction in Bert-Based Models for Reviewer System," no. May, 2023.

[12] A. Yovi, S. Adi, E. Alexander, U. Katolik, D. Cendika, and K. Ngasem, "Merangkum Teks Word Dan Pdf," vol. 12, no. 1, 2024.

[13] Nabil Haidarrahman Pribadi, "Sistem Rekomendasi Karya IlmiahBerdasarkan Semantic SimilarityMenggunakan Fasttext Dan MetodeWord Mover'S Distance," *Insitur Teknol. Sepuluh Nop.*, 2020.

[14] MEILINAEKA, "Web Scraping : Pengertian dan Fungsinya dalam Pengambilan Data." [Online]. Available: https://it.telkomuniversity.ac.id/web-scraping-pengertian-dan-fungsinya-dalam-pengambilan-data/

[15] J. Pragantha, T. Informatika, F. T. Informasi, and U. Tarumanagara, "Automatic Summarization Pada," vol. 1, no. 1, pp. 71–78, 2017.

[16] D. Darmanto, N. I. Pradasari, and E. Wahyudi, "Sistem Deteksi Plagiarisme Tugas Akhir Mahasiswa Berbasis Natural Language Processing Menggunakan Algoritma Jaro-Winkler dan TF-IDF," Smart Comp: Jurnalnya Orang Pintar Komputer, vol. 13, no. 1, pp. 201–211, 2024.

[17] A. Hexahost, "Text Preprocessing: Apa itu, Mengapa Penting, dan Bagaimana Melakukannya?," Hexahost. [Online]. Available: https://hexahost.id/pengertian-text-preprocessing/

[18] D. Septiani and I. Isabela, "Analisis Term Frequency Inverse Document Frequency (TF-IDF) Dalam Temu Kembali Informasi Pada Dokumen Teks," *SINTESIA J. Sist. dan Teknol. Inf. Indones.*, vol. 1, no. 2, pp. 81–88, 2023.

[19] "Different Techniques for Sentence Semantic Similarity in NLP," GeeksForGeeks. [Online]. Available: https://www.geeksforgeeks.org/different-techniques-for-sentence-semantic-similarity-in-nlp/