Journal Global Tecnology Computer
Vol 3, No. 3, Agustus 2024, Hal 85-94
ISSN: 2809-6118 (Online - Elektronik)
DOI: 10.47065/jogtc.v3i3.6256
Website https://ejurnal.seminar-id.com/index.php/jogtc

Natural Language Processing Ekstraksi Akronim Dan Ekspansi Pada Artikel Berbahasa Indonesia Menggunakan Metode Text Mining Dan Term Frequency-Inverse Document Frequency

Bahrus Sobri Pulungan

Fakultas Ilmu Komputer dan Teknologi Informasi, Program Studi Teknik Informasi, Universitas Budi Darma, Medan, Indonesia
Jl. Sisingamangaraja No.338, Siti Rejo I, Kec. Medan Kota, Kota Medan, Sumatera Utara, Indonesia
Email: sobriaweng12@gmail.com

Abstrak-Akronim adalah singkatan dari gabungan beberapa huruf atau suku kata ditulis dan diucapkan sebagai kata-kata sesuai dengan aturan fonologis bahasa terpengaruh. Kepanjangan dari sebuah akronim disebut ekspansi. Ekstraksi akronim dan ekspansi merupakan salah satu tugas *text mining* dalam bidang information retrieval yang digunakan pada *search engine*. *Search engine* membutuhkan database akronim dan ekspansi dalam menentukan hasil pencarian informasi yang relevan.Permasalahan yaitu sering terjadi ketika seseorang ataupun peneliti membuat suatu karya ilmiah pada terkhususnya penelitian di Indonesia yang mengabaikan ekstraksi akronim dari setiap kata yang digunakan ataupun kurang tepat, sehingga diperlukan suatu cara agar untuk mengatasi hal tersebut untuk membuat aplikasi ataupun media untu mendeteksi ekstraksi akronim tersebut dengan menerapkan Algoritma *Text Mining* dan *Term Frequency-Inverse Document Frequency* (TF-IDF). Berdasarkan permasalahan yang terdapat pada penelitian ini, maka penulis tertarik dalam melakukan penelitian pada skripsi dengan judul "*Natural Language Processing Ekstraksi* Akronim Dan Ekspansi Pada Artikel Bahasa Indonesia Menggunakan Metode *Text Mining* Dan *Term Frequency-Inverse Document Frequency* (TF-IDF)". Berdasarkan hasil perhitungan dengan TF-IDF, pada ekstraksi akronim dan ekspansi adalah nilai bobot yang didapat yaitu dengan nilai bobot -0,053 berdasarkan hal tersebut maka memproleh kalimat yang diekstraksi tersebut.

Kata Kunci: Ekstraksi; Akronim; Ekspansi; Artikel; Bahasa Indonesia.

Abstract-*Acronyms are abbreviations of combinations of several letters or syllables written and pronounced as words according to the phonological rules of the affected language. The extension of an acronym is called expansion. Acronym extraction and expansion is one of the text mining tasks in the field of information retrieval used in search engines. Search engines require a database of acronyms and expansion in determining search results for relevant information. The problem is that it often occurs when someone or a researcher makes a scientific work, especially research in Indonesia, which ignores the extraction of acronyms from each word used or is not quite right, so a way is needed to overcome this by creating an application or media to detect the extraction of the acronym using applying the Text Mining Algorithm and Term Frequency-Inverse Document Frequency (TF-IDF). Based on the problems contained in this research, the author is interested in conducting research on a thesis with the title "Natural Language Processing Acronym Extraction and Expansion in Indonesian Articles Using Text Mining Methods and Term Frequency-Inverse Document Frequency (TF-IDF)". Based on the results of calculations with TF-IDF, in acronym extraction and expansion, the weight value obtained is with a weight value of -0.053. Based on this, the extracted sentence is obtained.*

Keywords: *Extraction; Acronyms; Expansion; Articles; Indonesian.*

1. PENDAHULUAN

Akronim adalah singkatan dari gabungan beberapa suku kata atau huruf dapat dieja dan diucapkan sebagai kata-kata sesuai dengan aturan fonologis bahasa terpengaruh. Kepanjangan dari sebuah akronim disebut ekspansi. Ekstraksi adalah proses memisahkan informasi atau data yang relevan dari sumber yang lebih besar atau kompleks sering digunakan dalam berbagai konteks, termasuk ilmu data, pemprosesan bahasa alami, dan bidang lainnya. Penggunaan akronim dan ekspansi sering ditemui dalam sebuah tulisan, salah satunya pada artikel. Sebuah Artikel adalah sebuah komposisi tulisan yang berisi informasi atau penjelasan tentang suatu topik tertentu, artikel dapat ditulis dalam berbagai bahasa, termasuk bahasa indonesia. akronim digunakan saat ekspansi telah dijelaskan sebelumnya dalam artikel pada awal kalimat, misalnya Bahrus Sobri Pulungan (BSP).

Ekstraksi akronim dan ekspansi merupakan salah satu tugas *text mining* dalam bidang information retrieval yang digunakan pada *search engine*. *Search engine* membutuhkan database akronim dan ekspansi dalam menentukan hasil pencarian informasi yang relevan. Hal ini dapat diartikan sebagai contoh, ketika pengguna menginputkan sebuah *query* pada *search engine*, maka dokumen yang mengandung Bahrus Sobri Pulungan dan BSP harus diperhitungkan pada search engine, karena BSP merupakan akronim dari Bahrus Sobri Pulungan. Sehingga kemampuan search pada engine yang Bahrus Sobri Pulungan dapat mengganti akronim dan ekspansi begitu pula sebaliknya inilah yang membuat hasil pencarian informasi menjadi relevan. Akan tetapi, beberapa akronim memiliki lebih dari satu ekspansi yang dapat menyebabkan informasi yang didapatkan menjadi tidak optimal, jika akronim tersebut tidak diketahui kepanjangannya karena beberapa akronim memiliki singkatan yang sama, namun kepanjangan yang berbeda. Bahasa Indonesia adalah bahasa resmi dan bahasa persatuan di Republik Indonesia, dan akronim adalah singkatan yang terdiri dari gabungan huruf atau suku kata yang dilafalkan sebagai kata yang lengkap. Akronim sering digunakan untuk menghemat ruang atau waktu dalam komunikasi [1].

Maka penelitian ini mempunyai suatu permasalahan yaitu sering terjadi ketika seseorang ataupun peneliti membuat suatu karya ilmiah pada terkhususnya penelitian di Indonesia yang mengabaikan ekstraksi akronim dari setiap kata yang digunakan ataupun kurang tepat, sehingga diperlukan suatu cara agar untuk mengatasi hal tersebut untuk membuat

Journal Global Tecnology Computer
Vol 3, No. 3, Agustus 2024, Hal 85-94
ISSN: 2809-6118 (Online - Elektronik)
DOI: 10.47065/jogtc.v3i3.6256
Website https://ejurnal.seminar-id.com/index.php/jogtc

aplikasi ataupun media yang menggunakan Algoritma Text Mining dan *Term Frequency-Inverse Document Frequency* **(**TF- IDF) untuk mendeteksi akronim dalam teks.

Text Mining adalah proses analisis teks yang bertujuan untuk mencari pola, meringkas teks, dan menggali informasi yang relevan dari dokumen atau sumber teks lainnya. Ini digunakan untuk meningkatkan pemahaman atas konten teks yang ada, memudahkan pembaca dalam memahami konten, menghilangkan kebingungan yang mungkin terjadi, dan dapat mengidentifikasi kata-kata yang paling mewakili isi tersebut. TF-IDF *(Term Frequency-Inverse Document Frequency)* ini adalah metode yang digunakan dalam Text Mining untuk menghitung bobot kata-kata dalam dokumen dengan mempertimbangkan seberapa sering kata tersebut muncul dalam dokumen tertentu dan seberapa umum kata tersebut dalam seluruh koleksi dokumen.

Menurut penelitian terdahulu yang dilakukan oleh Meylita Putri Simatupang, Dito Putro Utomo dalam penelitiannya berjudul mengenai "Analisa Testimonial Dengan Menggunakan Algoritma Text Mining Dan Term Frequency- Inverse Document Frequence (TF-IDF) Pada Toko Allmeeart", hasil yang didapatkan dengan menerapkan algoritma text mining dan pengguna term frequency – inverse document frequency (TF-IDF) memang dapat digunakan untuk melakukan klasifikasi testimonial atau ulasan konotasi positif dan negative [2].

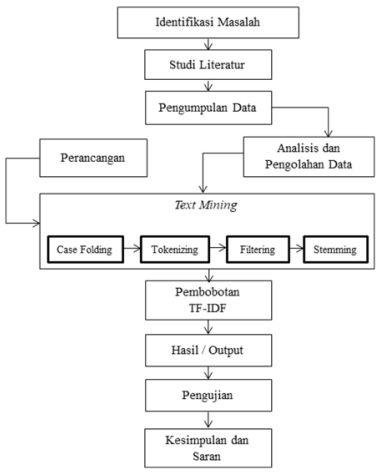
Menurut penelitian terdahulu yang dilakukan oleh Tri Putri Lestari dalam penelitiannya berjudul "Analisis Text Mining pada Sosial Media Twitter Menggunakan Metode Support Vector Machine dan Social Network Analysis", hasil penelitian yang menunjukkan bahwa metode Support Vector Machine (SVM) yang digunakan untuk mengklasifikasi pinjaman online dengan tweet berbahasa Indonesia menggunakan ekstraksi fitur TF-IDF memiliki tingkat akurasi yang baik sebesar 86.6%. Selain itu, Precision (presisi) yang tinggi pada kategori Positif dan Negatif menunjukkan kemampuan SVM dalam mengklasifikasikan tweet dengan konotasi positif dan negatif dengan akurasi yang tinggi. Namun, F1-Score yang lebih rendah pada kategori Netral mungkin menunjukkan adanya tantangan dalam mengklasifikasikan tweet sebagai Netral [3].

Menurut penelitian terdahulu yang dilakukan oleh Agung Daniel Sipayung, Fauziah, Nurhayati dalam penelitiannya berjudul mengenai "Sistem Aplikasi Penilaian Jawaban Essay Test Calon Karyawan PT Siloam Hospitals TB Simatupang Menggunakan Algoritma Text Mining TF-IDF Berbasis Web", hasil yang didapatkan dengan implementasi aplikasi penilaian jawaban ujian essay menggunakan Algoritma Text Mining TF-IDF. Dengan menggunakan teknologi web berbasis PHP, MySQL, dan algoritma TF-IDF, Anda dapat memproses jawaban ujian essay secara efisien dan memberikan hasil penilaian yang akurat. Selain itu, integrasi dengan use case diagram, class diagram, sequence diagram, statechart diagram, dan activity diagram memungkinkan pemahaman yang jelas tentang desain sistem. Penerapan aplikasi ini akan mengurangi pekerjaan manual yang harus dilakukan oleh HRD atau Recruitment Officer, yang akan menghemat waktu dan sumber daya, serta meminimalkan potensi kesalahan manusia dalam proses penilaian. [4].

Menurut penelitian terdahulu yang dilakukan oleh Indra Mawanta, T S Gunawan, Wanayumini dalam penelitiannya berjudul "Kemiripan Kalimat Judul Tugas Akhir dengan Metode Cosine Similarity dan Pembobotan TF-IDF", hasil yang didapatkan dengan penerapan pembobotan TF-IDF pada kalimat judul laporan tugas akhir dengan waktu efektif sekitar 0.12117 menit adalah langkah yang efisien untuk membantu dalam menentukan apakah judul laporan tugas akhir tersebut memenuhi kriteria atau tidak. Dengan metode ini, prodi atau program studi dapat lebih efisien dalam proses pengesahan judul dan mengurangi kemiripan isi dari laporan tugas akhir mahasiswa.TF-IDF membantu dalam mengidentifikasi kata-kata kunci yang paling penting dalam judul laporan tugas akhir dan memberikan bobot pada kata-kata tersebut berdasarkan frekuensi kemunculan mereka dalam dokumen [5].

2. METODOLOGI PENELITIAN

2.1 Kerangka Penelitian.



Gambar 1. Kerangka Kerja Penelitian

Vol 3, No. 3, Agustus 2024, Hal 85-94
ISSN: 2809-6118 (Online - Elektronik)
DOI: 10.47065/jogtc.v3i3.6256
Website https://ejurnal.seminar-id.com/index.php/jogtc

Berikut ini penjelasan dari tahapan – tahapan kerangka penelitian yang ada pada penelitian ini:

a. Tahap Identifikasi Masalah
   Penulis menguraikan apa yang menjadi sumber masalah pada ekstraksi akronim dan ekspansi pada artikel bahasa indonesia.
b. Tahap Studi Literatur
   Penulis mempelajari dan memahami teori-teori yang menjadi pedoman dan referensi yang diperoleh dari berbagai buku, jurnal, situs yang relevan terkait akronim dan ekspansi, algoritma atau metode yang digunakan, serta teori- teori yang menunjang materi penelitian.
c. Tahap Pengumpulan Data
   Data yang digunakan yaitu data pada artikel berbahasa indonesia mengenai ekstraksi akronim dan ekspansi.
d. Tahap Analisa dan Pengolahan Data
   Tahap analisa dan pengolahan data, diawali dengan menganalisa cara penggunaan, skema kerja dan bahasa pemograman yang dipakai pada aplikasi *Visual Basic.Net*, selanjutnya melakukakan pengolahan pada ekstraksi akronim dan ekspansi menggunakan metode *text mining* dan TF-IDF. Data yang diolah dengan tujuan agar mendapatkan beberapa informasi terkait ekstraksi akronim dan ekspansi pada artikel berbahasa indonesia.
e. Tahap Perancangan
   Tahap perancangan ini memberikan gambaran mengenai perancangan yang datanya telah di olah kedalam bentuk yang sederhana, mudah dan dapat dimengerti oleh *user.*
f. Tahap *Text mining*
   Tahap ini dilakukan persiapan teks yang belum terstruktur menjadi data yang baik dan siap untuk diolah, tidak terdapat aturan baku terkait tahapan dalam melakukan *text mining.* Proses yang dilakukan *text mining* meliputi tahapan *case folding, tokenizing, filtering* dan *steeming.*
g. Tahap Pembobotan TF – IDF
   Tahap pembobotan TF – IDF Digunakan pada penelitian ini untuk memilih fitur sebagai hasil ringkasan dengan penerapannya pada seleksi fitur bobot kata.
h. Tahap Hasil/*Output*
   Tahap ini merupakan proses pemeringkatan dokumen berdasarkan relevansi terhadap kueri pencarian, kata kata yang memiliki yang memiliki skor tinggi dalam sebuah dokumen menghasilkan tingkat yang lebih tinggi dalam konteks dokumen tersebut.
i. Tahap Pengujian
   Tahap pengujian, dilakukan pengujian untuk mengukur keakuratan dan kemampuan aplikasi yang dibangun dan mendapatkan hasil yang signifikan terhadap ekstraksi akronim dan ekspansi yang telah selesai diuji dan dapat menyelesaikan masalah pengguna.
j. Kesimpulan dan Saran
   Kesimpulan dan saran memuat hasil dari ekstraksi akronim dan ekspansi menggunakan metode *text mining* dan TF–IDF dan hal–hal yang perlu diperbaiki.

2.2 *Natural Language Processing* (NLP).

Natural Language Processing (NLP) merupakan pengembangan program yang memiliki kemampuan untuk memahami bahasa manusia. Prinsip dasar dari bahasa alami adalah sebagai representasi pesan yang ingin disampaikan antara individu. Bentuk utamanya mungkin berupa ucapan atau tulisan. Melalui pengolahan bahasa alami, diharapkan pengguna dapat berinteraksi dengan komputer menggunakan bahasa sehari-hari. Tujuannya adalah menciptakan model komputasi berdasarkan bahasa agar terjadi interaksi antara manusia dan komputer melalui bahasa alami [4].

2.3 Ekstraksi akronim.

Ekstraksi akronim adalah proses identifikasi dan pengambilan akronim dari teks atau dokumen. Ini melibatkan pengenalan kata-kata atau frasa yang merupakan singkatan dari frasa panjang atau istilah tertentu. Proses ini dapat digunakan untuk mengidentifikasi dan memahami makna akronim dalam konteks tertentu, yang seringkali berguna dalam analisis teks dan pemrosesan bahasa alami, yang diucapkan sebagai sebuah kata, akronim adalah salah satu bentuk dari singkatan, yang merupakan proses pemendekan kata atau suku kata yang diambil dari setiap kata dan diucapkan sebagai satu kata.

Salah satu hal yang membedakan akronim dari singkatan lainnya adalah bahwa akronim biasanya diucapkan sebagai sebuah kata tunggal, seperti "NATO" yang diucapkan sebagai "nay-toh". Dalam akronim, huruf awal dari setiap kata dalam frasa aslinya digunakan untuk membentuk kata baru yang mewakili keseluruhan frasa tersebutdengan inisial FBI misalnya, yang diucapkan dengan mengucapkan setiap huruf secara terpisah. Tapi, kebanyakan orang mengabaikan perbedaan tersebut. Dalam gaya penulisan ilmiah akronim tidak saja mencakup akronim yang telah disebutkan diatas, namun juga mencakup singkatan yang digunakan selain huruf pertama dari sebuah kata (seperti nm untuk "nanometer" atau Mr. untuk "mister"). Di sini, "akronim" akan digunakan secara bebas untuk singkatan apa pun [5].

2.4 Ekspansi.

Ekspansi adalah salah satu tugas yang penting dalam bidang *Natural Language Processing, text mining, dan information*

Journal Global Tecnology Computer
Vol 3, No. 3, Agustus 2024, Hal 85-94
ISSN: 2809-6118 (Online - Elektronik)
DOI: 10.47065/jogtc.v3i3.6256
Website https://ejurnal.seminar-id.com/index.php/jogtc

retrieval. Umumnya, ekstraksi akronim dan ekspansi dilakukan dengan dua pendekatan yaitu pendekatan heuristic yang melibatkan NLP dan pendekatan *pattern rule* atau *rule based*, dan pendekatan *machine learning*. Pendekatan NLP umumnya menggunakan teknik seperti tag *part of speech* (POS) dan relasi dari ekstraksi untuk menemukan ekspansi dari akronim. Kemudian *rule-based* dan *regular expression* digunakan selama proses ekstraksi akronim dan ekspansi yang dilakukan pada pendekatan berdasarkan pattern.

Aturan-aturan tersebut secara manual ditulis dengan mempertimbangkan karakter masing-masing dari beberapa tipe akronim, seperti ambigu, huruf kapital, panjang kata. Akronim sejenis singkatan yang disusun dari huruf pertama dari kata-kata dalam frase. Ini juga disebut sebagai deskriptor singkat frase. Akronim disebut sebagai kata, definisinya disebut sebagai ekspansi [6].

2.5 *Text Mining.*

Algoritma *text mining* memiliki peran penting dalam mengubah koleksi teks menjadi representasi numerik sehingga dapat diolah oleh komputer. *Text mining* merupakan subbidang khusus dari data mining yang bertujuan untuk menggali informasi dari teks. Menurut buku "*The Text Mining Handbook," text mining* dapat didefinisikan sebagai proses di mana pengguna berinteraksi dengan koleksi dokumen menggunakan alat analisis, yang merupakan komponen- komponen dari data mining, termasuk peringkatan dokumen. *Text mining* sering digunakan dalam klasifikasi dokumen teks, di mana dokumen-dokumen tersebut diberi label berdasarkan topik atau kategori mereka.

Dengan bantuan *text mining*, kita dapat mengenali jenis atau kategori dari sebuah artikel melalui kata-kata yang terdapat di dalamnya. Ini memungkinkan pengelompokkan dokumen yang efisien dalam waktu singkat. Dengan demikian, *text mining* membantu memahami dan mengelola informasi dari teks dengan lebih efektif, memungkinkan penggunaan yang luas dalam berbagai konteks, termasuk pengelompokkan dokumen dan analisis topik [7]. Ada sejumlah tahapan dalam menerapkan text mining. Berikut adalah langkah-langkah dalam *text mining* [8]:

a. *Tokenizing* yang merupakan tahapan awal dalam mengurai deskripsi yang berbentuk kalimat menjadi potongan-potongan kata. Berikut ini adalah proses tokenisasi pada beberapa kalimat.
b. *Filtering* adalah tahap proses dimana kita menghapus kata-kata yang tidak relevan (stoplist) atau menyimpan kata-kata yang penting (wordlist). Stoplist, atau biasa disebut stopwords, adalah kumpulan kata-kata yang tidak memiliki nilai deskriptif dan dapat diabaikan dalam pendekatan bag-of-words. Contoh kata-kata stopwords termasuk *"yang," "dan," "di," "dari,"* dan sebagainya. Daftar stopwords biasanya diperoleh dari kamus kata-kata yang tidak relevan atau disebut kamus tala, yang digunakan untuk menghilangkan kata-kata yang tidak penting dalam analisis teks. Berikut proses *filtering* pada tabel dibawah ini.
c. *Stemming* adalah langkah dalam text mining yang melibatkan pengubahan kata-kata yang memiliki imbuhan menjadi bentuk dasarnya atau kata dasar. Singkatnya, stemming bertujuan untuk mengembalikan kata-kata ke bentuk asalnya.
d. *Tagging* adalah tahap yang digunakan untuk mengidentifikasi bentuk dasar atau kata root dari setiap kata dalam teks, terutama setelah proses stemming. Tagging berguna untuk mengonversi kata dasar ke bentuk lampau. Proses ini umumnya digunakan dalam bahasa Inggris untuk mengembalikan kata-kata ke bentuk mereka yang benar.

2.6 *Term Frequency-Inverse Document Frequency* (TF-IDF).

Algoritma TF-IDF adalah pendekatan untuk memberikan bobot pada kata- kata dalam sebuah dokumen. Metode ini menggabungkan dua konsep, yaitu frekuensi kemunculan kata dalam dokumen tertentu (Term Frequency - TF) dan invers inversi frekuensi dokumen yang mengandung kata tersebut (Inverse Document Frequency - IDF). Frekuensi kata dalam dokumen menunjukkan seberapa penting kata itu dalam dokumen tersebut. Sehingga, kata akan memiliki bobot yang tinggi dalam hubungannya dengan dokumen jika frekuensinya tinggi dalam dokumen itu sendiri, namun rendah dalam frekuensinya di seluruh kumpulan dokumen. TF-IDF sering digunakan dalam penelusuran informasi dan text mining untuk memberikan bobot pada kata-kata yang relevan dalam teks [8].

2.6.1 *Term frequency.*

Term frequency (TF) adalah cara untuk mengukur seberapa sering suatu kata muncul dalam sebuah dokumen atau dalam koleksi dokumen tertentu. Dalam konteks pemrosesan bahasa alami dan analisis teks, term frequency (TF) digunakan untuk mengukur seberapa sering suatu kata muncul dalam sebuah teks atau dokumen. Perhitungan TF biasanya dilakukan dengan membagi jumlah kemunculan kata tertentu dalam teks dengan jumlah total kata dalam teks tersebut. Dengan kata lain, TF adalah rasio antara jumlah kemunculan kata dengan jumlah total kata dalam teks [8]. Cara kerja dalam mencari nilai term-frequency (Tf) dapat dihitung menggunakan persamaan berikut:

$$Tf_{t,d} = 1 + {}^{10}\text{Log } tf \ldots \quad (1)$$

keterangan dalam persamaan tersebut:
tf : *term frequency*, yaitu banyaknya kata atau term dalam dokumen.
$Tf_{t,d}$: nilai *term frequency* atau banyaknya kata tertentu $_t$ dalam dokumen $_d$ ini juga disebut pembobotan lokal.

2.6.2 *Inverse Document Frequency*

Inverse Document Frequency (IDF) adalah metode yang digunakan untuk mengukur seberapa penting suatu kata dalam koleksi dokumen secara keseluruhan. IDF digunakan dalam kombinasi dengan menghitung seberapa sering kata tersebut

Journal Global Tecnology Computer
Vol 3, No. 3, Agustus 2024, Hal 85-94
ISSN: 2809-6118 (Online - Elektronik)
DOI: 10.47065/jogtc.v3i3.6256
Website https://ejurnal.seminar-id.com/index.php/jogtc

muncul dalam seluruh koleksi dokumen. Konsep IDF mengasumsikan bahwa kata-kata yang muncul lebih jarang dalam koleksi dokumen umumnya lebih informatif daripada kata-kata yang muncul secara umum. IDF dapat membantu mengidentifikasi kata-kata kunci yang dapat membedakan atau memberikan informasi yang lebih berharga dalam sebuah dokumen [8]. Rumus untuk menghitung *inverse document frequency* (IDF) dalam persamaan 2:

$$Idf_t = {}^{10}\log {}^{n}/df_t \quad (2)$$

keterangan dari variabel sebagai berikut:
Idf_t : nilai *inverse document-frequency* atau pembobotan global untuk kata (term) tertentu.
^{n} : jumlah total dokumen dalam koleksi dokumen.
df_t : jumlah dokumen yang memiliki kata $_t$ tersebut.

Berdasarkan persamaan 1 (Term Frequency - TF) dan persamaan 2 (Inverse Document Frequency - IDF) dengan benar untuk menghitung bobot term ($W_{t,d}$) dalam dokumen (d). Persamaan 3 sebagai berikut:

$$W_{t,d} = tf_{t,d}.idf_t \quad (3)$$

keterangan:
$Tf_{t,d}$: jumlah kemunculan *term* $_t$ dalam dokumen $_d$
idf_t : nilai inverse document frequency (IDF) untuk kata (term) $_t$ tersebut
$W_{t,d}$: bobot *term* $_t$ terhadap dokumen $_d$

3. HASIL DAN PEMBAHASAN

3.1 Sampel data.

Sampel data yang digunakan penulis diambil dari artikel berbahasa Indonesia pada tabel 1 ini adalah contoh kalimat akronim dan ekpansi. Sampel data dapat dilihat pada tabel seperti dibawah:

Tabel 1. Sampel Data

No.	Artikel bahasa Indonesia	Akronim	Ekspansi
1.	Petani yang terlibat dalam usaha perkebunan kelapa sawit umumnya memiliki fokus pada pencapaian produksi yang maksimal. Salah satu komoditas utama dalam usaha perkebunan kelapa sawit adalah Tandan Buah Segar (TBS) yang kemudian dijual ke pabrik untuk diolah menjadi Crude Palm Oil (minyak kelapa sawit mentah). Produksi kelapa sawit adalah kegiatan yang melibatkan perawatan tanaman kelapa sawit itu sendiri. Ketika tanaman kelapa sawit dikelola dengan baik, hasil produksi TBS juga cenderung menjadi baik dan berkualitas.	TBS	Tandan Buah Segar
2.	Kemampuan mempertahankan ingatan adalah salah satu bagian penting dari keterampilan belajar dalam konteks proses pengajaran dan pembelajaran (PDP).	PDP	Pengajaran Dan Pembelajaran
3.	Dalam sektor pelayanan jasa, peran seorang notaris adalah sebagai pejabat umum yang memiliki kewenangan yang diberikan oleh negara untuk memberikan layanan kepada masyarakat, terutama dalam konteks pembuatan akta autentik. Hal ini diatur dalam Pasal 1868 Kitab Undang-undang Hukum Perdata (KUHPer).	KUHPer	Kitab Undang-undang Hukum Perdata
4.	Tindak pidana merujuk pada perbuatan yang dilarang dan dikenai sanksi pidana sesuai dengan ketentuan undang-undang. Aturan hukum materiil terkait tindak pidana telah diatur secara tertulis dan dikodifikasi dalam Undang- Undang Nomor 1 Tahun 1946 tentang Peraturan Hukum Pidana, yang umumnya dikenal dengan sebutan Kitab Undang-undang Hukum Pidana (KUHP).	KUHP	Kitab Undang-undang Hukum Pidana
5.	Sebelum memasuki proses pemeriksaan di hadapan pengadilan, pihak kepolisian membuat Berita Acara Pemeriksaan (BAP) yang berisi catatan tentang keterangan-keterangan yang diberikan oleh tersangka dan saksi-saksi selama proses penyelidikan atau pemeriksaan awal.	BAP	Berita Acara Pemeriksaan
6.	Energi adalah salah satu kebutuhan pokok manusia yang terus meningkat seiring dengan peningkatan taraf hidup. Bahan Bakar Minyak (BBM) memiliki peran yang sangat dominan dalam memenuhi kebutuhan energi nasional.	BBM	Bahan Bakar Minyak
7.	Adanya perkembangan Ilmu Pengetahuan Teknologi (IPTEK) yang pesat berdampak pada perilaku manusia dalam kehidupan sosial dan negara, menjadikannya semakin kompleks.	IPTEK	Ilmu Pengetahuan Teknologi
8.	Pelayanan tes Surat Izin Mengemudi (SIM) merujuk pada pelayanan SIM yang dilakukan sesuai dengan prosedur dan ketentuan yang telah diatur dalam Undang-Undang.	SIM	Surat Izin Mengemudi
9.	Kartu Keluarga Sejahtera (KKS) memiliki dua fungsi utama, yaitu sebagai indikator status masyarakat yang kurang mampu dan sebagai kartu identitas untuk mendapatkan manfaat dari program Simpan Keluarga Sejahtera.	KKS	Kartu Keluarga Sejahtera
10.	Seperti, Perlindungan tersebut melibatkan pelayanan negara yang mencakup penerbitan dokumen kependudukan, seperti Nomor Induk Kependudukan (NIK). Peningkatan ketertiban dan integrasi administrasi kependudukan sangat penting untuk membantu dalam merumuskan kebijakan, merancang, dan melaksanakan berbagai program pembangunan dengan lebih efisien.	NIK	Nomor Induk Kependudukan

Journal Global Tecnology Computer
Vol 3, No. 3, Agustus 2024, Hal 85-94
ISSN: 2809-6118 (Online - Elektronik)
DOI: 10.47065/jogtc.v3i3.6256
Website https://ejurnal.seminar-id.com/index.php/jogtc

3.2 Penerapan Algoritma *Text Mining*

Berdasarkan penelitian ini, peneliti melakukan pengolahan data yang diambil dari bab 1 dan terhadap 10 sampel data berupa artikel berbahasa Indonesia yang akan diproses dengan menggunakan algoritma *text mining.* Adapun sampel data yang digunakan pada penelitian ini dapat dilihat pada table 1 sebagai berikut:

a. *Case Folding*
Case Folding adalah langkah dalam sistem untuk mengubah semua huruf besar dalam teks menjadi huruf kecil. Tahap case folding digunakan untuk mengkonversi seluruh teks menjadi format standar dan menghapus tanda baca serta karakter khusus. Hasil dari tahapan ini dapat dilihat dalam Tabel 2 di bawah ini:

Tabel 2. Proses *Case Folding*

Teks Input	Hasil *Case Folding*
Petani yang terlibat dalam usaha perkebunan kelapa sawit umumnya memiliki fokus pada pencapaian produksi yang maksimal. Salah satu komoditas utama dalam usaha perkebunan kelapa sawit adalah Tandan Buah Segar (TBS), yang kemudian dijual ke pabrik untuk diolah menjadi Crude Palm Oil (minyak kelapa sawit mentah). Produksi kelapa sawit adalah kegiatan yang melibatkan perawatan tanaman kelapa sawit itu sendiri. Ketika tanaman kelapa sawit dikelola dengan baik, hasil produksi TBS juga cenderung menjadi baik dan berkualitas.	petani yang terlibat dalam usaha perkebunan kelapa sawit umumnya memiliki fokus pada pencapaian produksi yang maksimal. salah satu komoditas utama dalam usaha perkebunan kelapa sawit adalah tandan buah segar (tbs), yang kemudian dijual ke pabrik untuk diolah menjadi crude palm oil (minyak kelapa sawit mentah). produksi kelapa sawit adalah kegiatan yang melibatkan perawatan tanaman kelapa sawit itu sendiri. ketika tanaman kelapa sawit dikelola dengan baik, hasil produksi tbs juga cenderung menjadi baik dan berkualitas.

b. *Tokenizing*
Tahap *tokenizing* merupakan proses mengurai kalimat menjadi kata-kata yang membentuknya. Berdasarkan tahap ini, kalimat yang diberikan oleh pengguna akan dipecah menjadi kalimat yang membentuknya. Tahap *tokenizing* dapat dilihat pada tabel 3 di bawah ini:

Tabel 3. Proses *Tokenizing*

Hasil *Case Folding*	Hasil *Tokenizing*
petani yang terlibat dalam usaha perkebunan kelapa sawit umumnya memiliki fokus pada pencapaian produksi yang maksimal. salah satu komoditas utama dalam usaha perkebunan kelapa sawit adalah tandan buah segar (tbs), yang kemudian dijual ke pabrik untuk diolah menjadi crude palm oil (minyak kelapa sawit mentah). produksi kelapa sawit adalah kegiatan yang melibatkan perawatan tanaman kelapa sawit itu sendiri. ketika tanaman kelapa sawit dikelola dengan baik, hasil produksi tbs juga cenderung menjadi baik dan berkualitas.	”petani”,”yang”,”terlibat”,”dalam”,”usaha”,”perkebunan”,”kelapa”,”sawit”,”umumnya”,”memiliki”,”fokus”,”pada”,”pencapaian”,”produksi”,”yang”,”maksimal”,”salah”,”satu”,”komoditas”,”utama”,”dalam”,”usaha”,”perkebunan”,”kelapa”,”sawit”,”adalah”,”tandan”,”buah”,”segar”,”yang”,”kemudian”,”dijual”,”ke”,”pabrik”,”untuk”,”diolah”,”menjadi”,”crude”,”palm”,”oil”,”minyak”,”kelapa”,”sawit”,”mentah”,”produksi”,”kelapa”,”sawit”,”adalah”,”kegiatan”,”yang”,”melibatkan”,”perawatan”,”tanaman”,”kelapa”,”sawit”,”itu”,”sendiri”,”ketika”,”tanaman”,”kelapa”,”sawit”,”dikelola”,”dengan”,”baik”,”hasil”,”produksi”,”tbs”,”juga”,”cenderung”,”menjadi”,”baik”,”dan”,”berkualitas”

c. *Filtering*
Berdasarkan pada tahap ini, dilakukan pengapusan kata yang tidak penting atau kata yang tidak digunakan dalam pemprosesan. Proses yang dilakukan pada tahap ini dibantu dengan memanfaatkan *library stopword* dari tala ini dibantu dengan penghapusan kata yang tidak relevan. Hasil dari tahapan ini dapat dilihat pada tabel 4 di bawah ini:

Tabel 4. Proses *Filtering*

Hasil *Tokenizing*	Hasil *Filtering*
”petani”,”yang”,”terlibat”,”dalam”,”usaha”,”perkebunan”,”kelapa”,”sawit”,”umumnya”,”memiliki”,”fokus”,”pada”,”pencapaian”,”produksi”,”yang”,”maksimal”,”salah”,”satu”,”komoditas”,”utama”,”dalam”,”usaha”,”perkebunan”,”kelapa”,”sawit”,”adalah”,”tandan”,”buah”,”segar”,”yang”,”kemudian”,”dijual”,”ke”,”pabrik”,”untuk”,”diolah”,”menjadi”,”crude”,”palm”,”oil”,”minyak”,”kelapa”,”sawit”,”mentah”,”produksi”,”kelapa”,”sawit”,”adalah”,”kegiatan”,”yang”,”melibatkan”,”perawatan”,”tanaman”,”kelapa”,”sawit”,”itu”,”sendiri”,”ketika”,”tanaman”,”kelapa”,”sawit”,”dikelola”,”dengan”,”baik”,”hasil”,”produksi”,”tbs”,”juga”,”cenderung”,”menjadi”,”baik”,”d an”,”berkualitas”	“petani”,”terlibat”,”usaha”,”perkebunan”,”kelapa”,”sawit”,”memiliki”,”fokus”,”pencapaia”,”produksi”,”maksimal”,”komoditas”,”usaha”,”perkebunan”,”kelapa”,”sawit”,”tandan”,”segar”,”dijual”,”pabrik”,”diolah”,”crude”,”palm”,”oil”,”minyak”,”kelapa”,”sawit”,”kegiatan”,”melibatkan”,”perawatan”,”tanaman”,”kelapa”,”sawit”,”dikelola”,”produksi”,”tbs”,”cendrun” ,”berkualitas”

Journal Global Tecnology Computer
Vol 3, No. 3, Agustus 2024, Hal 85-94
ISSN: 2809-6118 (Online - Elektronik)
DOI: 10.47065/jogtc.v3i3.6256
Website https://ejurnal.seminar-id.com/index.php/jogtc

d. *Stemming*
Tahap stemming merupakan kata dengan imbuhan akan diubah menjadi lebih bentuk dasar atau kata aslinya. Tujuan dari tahap ini untuk mengembalikan kata ke bentuk aslinya dengan cara menghilangkan imbuhan yang tidak diperlukan:

Tabel 5. Proses *Stemming*

Hasil *Filtering*	Hasil *Stemming*
“petani”,”terlibat”,”usaha”,”perkebunan”, ”kelapa”,”sawit”,”memiliki”,”fokus”,”pe ncapaia”,”produksi”,”maksimal”,”komod itas”,”usaha”,”perkebunan”,”kelapa”,”sa wit”,”tandan”,”segar”,”dijual”,”pabrik”,” diolah”,”crude”,”palm”,”oil”,”minyak”,” kelapa”,”sawit”,”kegiatan”,”melibatkan”, ”perawatan”,”tanaman”,”kelapa”,”sawit”, ”dikelola”,”produksi”,”tbs”,”cendrun”,”berkualitas”	“tani”,”libat”,”usaha”,”kebun”,”kelapa”,”sawit”,” milik”,”fokus”,”capai”,”produksi”,”maksimal”,”k omoditas”,”usaha”,”kebun”,”kelapa”,”sawit”,”tan dan”,”segar”,”jual”,”pabrik”,”olah”,”crude”,”pal m”,”oil”,”minyak”,”kelapa”,”sawit”,”giat”,”libat” ,”rawat”,”tanam”,”kelapa”,”sawit”,”kelola”,”prod uksi”,”tbs”,”cendrung”,”kualitas”

Setelah ekstraksi data akronim dan eksoansi diolah dengan algoritma *text mining* maka didapat hasil keseluruhan proses dari penerapan algoritma *text mining* seluruh hasil *text mining* dapat dilihat pada gambar 6 dan gambar 7 di bawah ini:

Tabel 6. Hasil Keseluruhan *Text Mining*

Hasil Keseluruhan *Text Mining*
“tani”,”libat”,”usaha”,”kebun”,”kelapa”,”sawit”,”milik”,”fokus”,”capai”,”produksi”,”maksimal”, ”komoditas”,”usaha”,”kebun”,”kelapa”,”sawit”,”tandan”,”segar”,”jual”,”pabrik”,”olah”,”crude”, ”palm”,”oil”,”minyak”,”kelapa”,”sawit”,”giat”,”libat”,”rawat”,”tanam”,”kelapa”,”sawit”,”kelola ”,”produksi”,”tbs”,”cendrung”,”kualitas”

Selanjutnya pada tabel 7 menunjukkan data pemrosesan text mining ekstraksi akronim dan expansi

Tabel 7. Data Pemprosesan *Text Mining* Ekstraksi Akronim dan Ekspansi

Data Akronim dan Ekspansi
“TBS: Tandan Buah Segar”,” PDP: Pengajaran Dan Pembelajaran”,”KUHPer: Kitab Undang- undang Hukum Perdata”,”KUHP: Kitab Undang-undang Hukum Pidana”,”BAP: Berita Acara Pemeriksaan”,”BBM: Bahan Bakar Minyak”,”IPTEK: Ilmu Pengetahuan Teknologi”,”SIM: Surat Izin Mengemudi”,”KKS: Kartu Keluarga Sejahtera”,”NIK: Nomor Induk Kependudukan”,”

3.3 Term Frequency-Inverse Document Frecuency (TF-IDF).

Langkah pada proses *text mining* menghasilkan kalimat penting untuk perhitungan, proses perhitungan TF-IDF dimulai. Mengetahui beberapa banyak kata yang muncul pada setiap dokumen, langkah selanjutnya adalah menghitung frekuensi kemunculan kata yang sama di seluruh dokumen. Setelah itu, dilakukan perhitungan *Inverse Document Frequency* (IDF), yang melibatkan pembagian jumlah dokumen dengan *Term Frequency* (TF). Jika hasil pembagian tersebut sama dengan jumlah dokumen (DF = N), maka nilai IDF adalah 0 (nol). Oleh karena itu, nilai 1 dapat ditambahkan ke nilai IDF. Langkah berikutnya adalah menghitung nilai Wdt, yang merupakan hasil perkalian antara nilai TF dan nilai IDF, menghasilkan bobot *term*. Anda dapat melihat hasilnya dalam Tabel 4.8 di bawah ini:

Tabel 8. Perhitungan TF-IDF

Term	TF	DF	IDF
	D1		Login(n/df)+1
Tani	1	1	1,000
libat	2	2	0,699
usaha	2	2	0,699
kebun	2	2	0,699
kelapa	4	4	0,398
sawit	4	4	0,398
milik	1	1	1,000
fokus	1	1	1,000
capai	1	1	1,000
produksi	2	2	0,699
maksimal	1	1	1,000
komoditas	1	1	1,000
tandan	1	1	1,000
segar	1	1	1,000
Jual	1	1	1,000
pabrik	1	1	1,000
Olah	1	1	1,000

Journal Global Tecnology Computer
Vol 3, No. 3, Agustus 2024, Hal 85-94
ISSN: 2809-6118 (Online - Elektronik)
DOI: 10.47065/jogtc.v3i3.6256
Website https://ejurnal.seminar-id.com/index.php/jogtc

Term	TF	DF	IDF
	D1		Login(n/df)+1
crude	1	1	1,000
palm	1	1	1,000
Oil	1	1	1,000
minyak	1	1	1,000
Giat	1	1	1,000
rawat	1	1	1,000
tanam	1	1	1,000
kelola	1	1	1,000
Tbs	1	1	1,000
cendrung	1	1	1,000
kualitas	1	1	1,000

Langkah berikutnya adalah menghitung Wdt, yang merupakan hasil perkalian antara nilai TF dan nilai IDF, sehingga menghasilkan bobot term. Proses pembobotan TF-IDF ini dapat dilihat dalam Tabel 4.10 di bawah ini

Tabel 9. Pembobotan TF-IDF

W=TF*IDF	
Term	**D1**
Tani	1,000
Libat	1,398
Usaha	1,398
Kebun	1,398
Kelapa	1,592
Sawit	1,592
Milik	1,000
Fokus	1,000
Capai	1,000
Produksi	1,398
maksimal	1,000
komoditas	1,000
tandan	1,000
segar	1,000
jual	1,000
pabrik	1,000
olah	1,000
crude	1,000
palm	1,000
oil	1,000
minyak	1,000
giat	1,000
rawat	1,000
tanam	1,000
kelola	1,000
tbs	1,000
cendrung	1,000
kualitas	1,000
Nilai	53,959

Berdasarkan pembobotan tersebut didapatkan hasil nilai bobot akronim dan ekspansinya yang dapat dilihat pada tabel 10 sebagai berikut:

Tabel 10. Nilai Bobot Ekstraksi Akronim Dan Ekspansi

Hasil Perhitungan				
Akronim	**Ekspansi**	**Hasil Ekstraksi**		
TBS	Tandan Buah Segar	TF (Term Frequency)	IDF (Inverse Document Frequency)	**Nilai Bobot TF- IDF**
5	15	-0,176	0,301	-0,053

Keterangan:
a. Jumlah teks Akronim"'tbs" dalam dokumen: 5
b. Jumlah kata Ekspansi "tandan"buah"segar" dalam dokumen: 15
c. TF (Term Frequency):
   1. jumlah kemunculan "tandan"buah"segar" dalam dokumen: 15

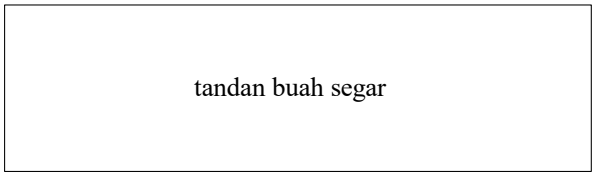
Journal Global Tecnology Computer
Vol 3, No. 3, Agustus 2024, Hal 85-94
ISSN: 2809-6118 (Online - Elektronik)
DOI: 10.47065/jogtc.v3i3.6256
Website https://ejurnal.seminar-id.com/index.php/jogtc

2. TF untuk “tandan”buah”segar”: 1/15 = -0,176

d. IDF (Inverse Document Frequency):
   1. IDF untuk “tbs”:log(5/1)= log(5) = 0,301

e. TF-IDF:
   1. Nilai TF-IDF untuk akronim dan ekspansinya dalam dokumen:
      -0,176*0,301 = -0,053

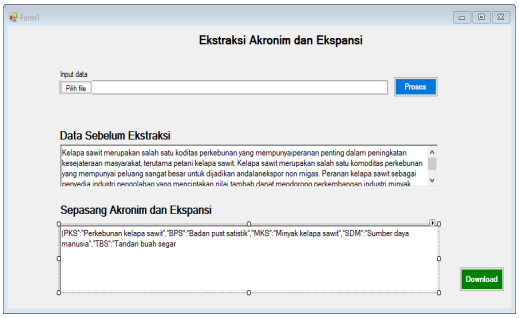
Berdasarkan hasil perhitungan dengan TF-IDF, pada ekstraksi akronim dan ekspansi adalah nilai bobot yang didapat yaitu dengan nilai bobot -0,053 berdasarkan hal tersebut maka memproleh kalimat yang diekstraksi tersebut, dapat dilihat pada Gambar 2 di bawah sebeagai berikut:



Gambar 2. Kalimat Yang Diekstraksi

3.4 Implementasi

Tahap implementasi adalah ketika sistem yang telah dikembangkan diuji coba. Bagian ini mencakup spesifikasi perangkat keras (*hardware*) dan perangkat lunak (*software*), serta hasil tampilan sistem ketika sedang berjalan. Tampilan halaman ini akan muncul ketika pengguna mengakases aplikasi ekstraksi akronim. Berikut ini tampilan dari aplikasi akronim dan ekspansi saat sudah dilakukan ekstraksi akronim dan ekspansi pada artikel berbahasa indonesia. Aplikasi ekstraksi akronim yang sudah dirancang dapat digunakan dengan baik dan hasil pengujian dapat dilihat pada gambar di bawah ini:



Gambar 3. Tampilan Aplikasi setelah Ekstraksi Akronim dan Ekspansi

4. KESIMPULAN

Berdasarkan analisis, tinjauan, dan evaluasi pada bab-bab sebelumnya, serta penelitian terhadap ekstraksi akronim dan ekspansi pada artikel berbahasa Indonesia, maka dapat diambil kesimpulan bahwa Dalam melakukan akronim dan ekspansi pada artikel berbahasa Indonesia menggunakan *text mining* dan tf-idf, dengan cara melakukan langkah-langkah seperti *case folding, tokenizing, filtering, dan stemming.* Agar dapat mengidentifikasi kata-kata yang paling penting. Hasil pada proses *text mining* menghasilkan kalimat penting untuk perhitungan, proses perhitungan TF-IDF dimulai. Mengetahui beberapa banyak kata yang muncul pada setiap dokumen, langkah berikutnya adalah menjumlahkan kata yang sama di semua dokumen. Berdasarkan hasil penerapan perhitungan dengan TF-IDF, pada ekstraksi akronim dan ekspansi adalah nilai bobot yang didapat yaitu dengan nilai bobot -0,053 berdasarkan hal tersebut maka memproleh kalimat yang diekstraksi tersebut. Penerapan ekstraksi akronim dan ekspansi dirancang menggunakan Microsoft *Visual Basic.Net* 2010 dengan menerapkan *text mining* dan tf-idf. Tujuan dari hal ini adalah untuk memberikan kemudahan kepada penulis dalam proses ekstraksi akronim dan ekspansi pada artikel berbahasa Indonesia.

REFERENCES

[1] A. F. Harahap and G. L. Ginting, “Penerapan Algoritma RAITA pada Kamus Akronim Bahasa Indonesia Berbasis Android,” *TIN Terap. Inform…*, vol. 1, no. 3, 2020, [Online]. Available: http://ejurnal.seminar-id.com/index.php/tin/article/view/426%0Ahttp://ejurnal.seminar- id.com/index.php/tin/article/download/426/276

[2] M. P. Simatupang and D. P. Utomo, “Analisa Testimonial Dengan Menggunakan Algoritma Text Mining Dan Term Frequency-Inverse Document Frequence (Tf-Idf) Pada Toko Allmeeart,” *KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer)*, vol. 3, no. 1, pp. 808–814, 2019, doi: 10.30865/komik.v3i1.1697.

[3] T. P. Lestari, “Analisis Text Mining pada Sosial Media Twitter Menggunakan Metode Support Vector Machine (SVM) dan Social Network Analysis (SNA),” *J. Inform. Ekon. Bisnis*, vol. 4, no. 3, pp. 65–71, 2022, doi: 10.37034/infeb.v4i3.146.

[4] R. Yusuf, T. A. Saputri, and A. A. Wicaksono, “Penerapan Natural Language Processing Berbasis Virtual Assistant Pada Bagian Administrasi Akademik Stmik Dharma Wacana,” *Int. Res. Big-Data Comput. Technol. I- Robot*, vol. 5, no. 1, pp. 33–47, 2022,

Vol 3, No. 3, Agustus 2024, Hal 85-94
ISSN: 2809-6118 (Online - Elektronik)
DOI: 10.47065/jogtc.v3i3.6256
Website https://ejurnal.seminar-id.com/index.php/jogtc

doi: 10.53514/ir.v5i1.228.
[5] Yosi Arisanti Linda, "104 | J u r n a l L I T E R A S I Volume 2 | Nomor 2 | Oktober 2018," *J u r n a l L I T E R A S I*, vol. 2, pp. 104–112, 2018.
[6] R. Menaha and V. E. Jayanthi, "A Survey on Acronym – Expansion Mining Approaches from Text and Web," 1921.
[7] R. A. Sasmita, A. Z. Falani, F. I. Komputer, U. N. Surabaya, and T. Mining, "Pemanfaatan algoritma tf/idf pada sistem informasi ecomplaint handling," vol. 27, no. 1, pp. 27–33, 2018.
[8] H. Sari, G. L. Ginting, and T. Zebua, "Penerapan Algoritma Text Mining dan TF-IDF Untuk Pengelompokan Topik Skripsi Pada Aplikasi Repository STMIK Budi Darma," vol. 2, no. 7, pp. 414–432, 2021.