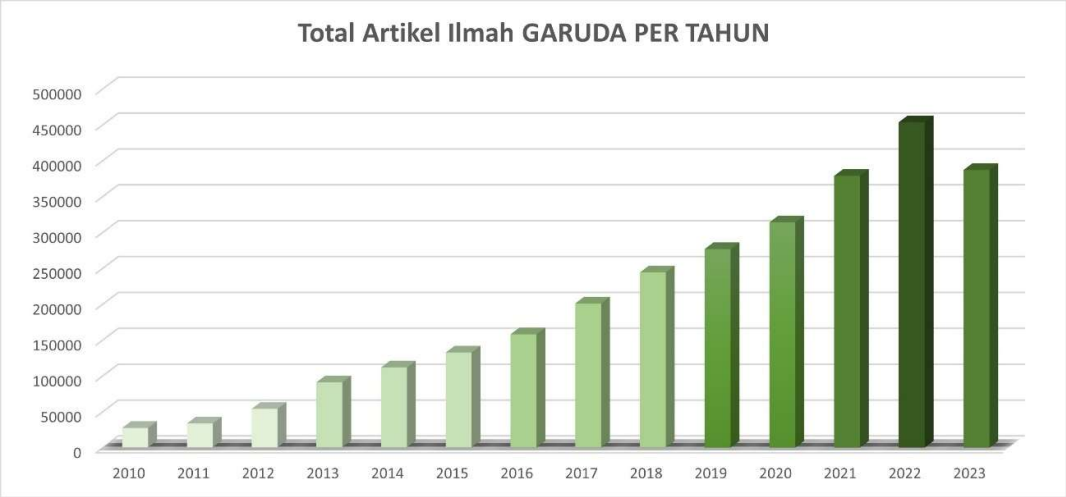
BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan zaman sekarang sudah banyak membawa perubahan di berbagai bidang, terutama dalam bidang teknologi khususnya akses informasi yang tersedia secara *up to date* (Pembangunan *et al*., n.d.). Teknologi informasi telah berkembang dalam dekade terakhir. Masyarakat disuguhkan dengan banyaknya informasi berita baik dari media cetak dan media daring atau *online* (Romadhoni, 2018). Untuk media cetak sering kita temui di majalah, buku, koran, dan lain-lain sedangkan media daring sering kita temui di internet contohnya seperti artikel yang bisa dilihat di *Website*. Artikel telah banyak membantu para peneliti dan juga pelajar untuk menemukan informasi yang sesuai dengan bidang mereka.



Gambar 1.1 Publikasi GARUDA 13 Tahun Terakhir

Menurut GARUDA (Garuda - Garba Rujukan Digital, 2024), terdapat tren peningkatan yang signifikan dalam jumlah artikel ilmiah yang terdaftar setiap tahunnya. Fenomena peningkatan ini menjadi perhatian penting dalam konteks pengembangan ilmu pengetahuan dan keilmuan. Data statistik yang mencerminkan pertumbuhan kuantitatif artikel pada Gambar 1.1. Grafik publikasi artikel ilmiah terus meningkat secara signifikan per tahunnya. Peningkatan ini mencerminkan dinamika dan relevansi peran artikel ilmiah sebagai sumber informasi dalam perkembangan

penelitian dan akademis. Seiring dengan tren peningkatan publikasi ilmiah, para peneliti dan pembaca memiliki akses yang lebih luas terhadap ilmu yang relevan dengan bidang studi yang mereka tekuni. Pembaca perlu menyerap informasi dengan efisien dan mengambil keputusan dengan cepat mengingat keterbatasan waktu yang dimiliki, sehingga membutuhkan ringkasan artikel yang cepat dan efisien (Pinto *et al*., n.d.). Untuk meresponi situasi ini, terdapat aplikasi peringkasan teks berbasis website untuk membantu peringkasan artikel secara cepat. Aplikasi berbasis website yang tersedia saat ini yaitu *paraphraser.io*, ringkasan yang dihasilkan aplikasi ini dengan membuat kalimat-kalimat baru yang berbeda dari teks sumber asli. Meskipun pendekatan ini fleksibel, namun berpotensi menyebabkan hilangnya informasi penting, distorsi makna, serta ketidakwajaran dalam kalimat yang dihasilkan (Paulus *et al*., 2017). Oleh karena itu, penelitian ini bertujuan untuk mengembangkan sistem peringkasan teks yang memiliki keunggulan dalam mempertahankan integritas informasi penting dari teks sumber dan menghindari distorsi makna dengan mengambil kalimat-kalimat secara harfiah. Sistem yang dikembangkan akan menerapkan metode pemilihan kalimat untuk menghasilkan ringkasan dari teks bahasa Indonesia yang panjang. Salah satu bidang ilmu teknologi yang berperan dalam pengelolaan informasi adalah *Natural Language Processing* (NLP).

Nature Language Processing (NLP) adalah subbidang kecerdasan buatan (*artificial intelligence*) yang berfokus pada memungkinkan komputer memahami dan memproses bahasa manusia (Zhou *et al*., 2020). NLP bertujuan untuk memungkinkan interaksi antara manusia dengan komputer melalui bahasa alami dan juga bertujuan untuk membantu komputer mengerti bahasa manusia secara lebih akurat dan memproses bahasa manusia dengan model komputasi melalui lisan maupun tulisan (Adjie Wicaksono *et al*., n.d.). Salah satu contoh teknik NLP adalah *Text summarization.*

Text summarization atau peringkasan teks adalah teknik dari NLP yang secara otomatis menghasilkan ringkasan yang berisi kalimat-kalimat penting dan mencakup semua informasi penting yang relevan dari dokumen asli (Widyassari *et al*., 2022). *Text summarization* bertujuan untuk menciptakan ringkasan teks dari dokumen asli yang koheren dari teks asli dengan menekankan dan mengidentifikasi poin-point penting dan memiliki Panjang teks yang lebih sedikit dibanding dokumentasi yang asli

(Babar & Patil, 2015). *Text summarization* terdiri dari dua pendekatan utama yakni *Abstractive Summarization* dan juga *extractive summarization. Abstractive summarization* melibatkan membuat ringkasan dalam bahasa alami baru, bukan dengan sekedar memilih beberapa kalimat dari dokumen sumber sedangkan *extractive summarization* melibatkan pemilihan kalimat-kalimat penting dari teks sumber dan menggabungkannya untuk membentuk ringkasan (Gupta & Lehal, 2010). Kalimat-kalimat yang dipilih tetap persis seperti dalam teks asli tanpa ada modifikasi. Dalam Penelitian ini , peneliti memilih *TextRank* dengan pembobotan *Term Frequency Inverse Document Frequency (*TF-IDF).

TextRank adalah algoritma peringkat berbasis grafik yang digunakan untuk menemukan kalimat dan kata kunci yang relevan (K *et al*., 2023). *TextRank* termasuk dalam *Unsupervised Learning* yaitu tanpa pembelajaran tanpa label (Mariani *et al*., 2023). *TextRank* merepresentasikan sebuah dokumen teks sebagai sebuah graf, di mana setiap kalimat dalam dokumen tersebut dianggap sebagai sebuah simpul (node). Hubungan antar node dalam kalimat ditentukan oleh nilai kemiripan (*similarity*) antara kalimat-kalimat tersebut. (Fadhila Hernawan *et al*., 2022). Salah satu keunggulan dalam *TextRank* yaitu tidak memerlukan proses pelatihan dengan menggunakan data *training* pada algoritma yang digunakan (Alfariz, 2022). Di sisi lain, *Term Frequency Inverse Document Frequency (*TF-IDF*)* adalah merupakan metode pembobotan yang mampu mengidentifikasi kata-kata penting dalam suatu dokumen dengan mempertimbangkan frekuensi kemunculan kata tersebut di dalam dokumen serta invers frekuensi kemunculannya di dalam keseluruhan dokumen (Komang *et al*., 2018). Metode ini menghitung frekuensi kemunculan kata dalam sebuah dokumen (TF) dan kemudian menyesuaikan dengan frekuensi kemunculan kata tersebut dalam seluruh dokumen (IDF) (Elektronik *et al*., n.d.). Sehingga frekuensi kemunculan kata dalam hubungan antara sebuah kata dan sebuah dokumen akan tinggi jika frekuensi kata tinggi di dalam dokumen (Sasmita & Falani, n.d.). Algoritma *TextRank* baik untuk menemukan kalimat penting, tetapi tidak memberikan bobot pada kata-katanya. *TextRank* sendiri hanya menggunakan kesamaan kalimat untuk membangun graf dan meringkas teks berdasarkan peringkat graf tersebut. Namun, dengan menggabungkannya dengan TF-IDF, pembobotan tambahan dapat diberikan pada kata-kata penting dalam setiap kalimat. Meskipun *TextRank* dapat diimplementasikan

tanpa bantuan algoritma lain, penggabungan dengan TF-IDF dilakukan untuk meningkatkan performa ekstraksi ringkasan. Hal ini disebabkan oleh kemampuan TF-IDF dalam memberikan pembobotan awal pada kata-kata penting yang berguna bagi *TextRank* untuk membangun graf yang lebih akurat, serta menghasilkan peringkat kalimat yang lebih relevan untuk disertakan dalam ringkasan akhir hasil dari penelitian tersebut akan di validasi oleh 1 orang ahli dari tiap bidang yaitu bidang bahasa dan bidang web.

Penelitian tentang *text summarization* pernah dilakukan oleh (Kodicherla *et al*., 2023) yaitu membandingkan kinerja dua metode yaitu *TextRank* dan *Latent Semantic Analysis* (LSA) untuk peringkasan berita. Dataset yang digunakan dalam penelitian ini berasal dari "BBC *News Summary Dataset*" yang terdiri dari 2225 dokumen dari situs web berita BBC yang mencakup artikel-artikel dari lima kategori berbeda: Bisnis, Politik, Hiburan, Teknologi, dan Olahraga. Hasil eksperimen menunjukkan bahwa *TextRank* memiliki kinerja yang lebih baik daripada LSA dalam hal *Precision*, *Recall*, dan *F-Score* untuk semua kategori berita. Perbandingan hasil evaluasi menggunakan metrik *ROUGE-N* dan *ROUGE-L* juga menunjukkan bahwa *TextRank* secara konsisten lebih baik dibandingkan LSA dalam hal nilai *Precision*, *Recall*, dan *F-Score*. Hal ini menunjukkan bahwa *TextRank* lebih mampu mengidentifikasi dan mengekstrak kalimat-kalimat penting dari dokumen asli untuk pembuatan ringkasan berita yang lebih baik. Penelitian tentang algoritma *TextRank* juga pernah dilakukan oleh (Christanti & Pragantha, 2017) yaitu membuat sistem *automatic summarization* bahasa Indonesia. Algoritma *TextRank* diterapkan untuk mendapatkan peringkat pada setiap kalimat. Kalimat dengan peringkat tertinggi kemudian dipilih untuk menjadi ringkasan ekstraktif dokumen sesuai persentase kompresi yang diinginkan (50% atau 75%). Hasil penelitian menunjukkan sistem dapat memberikan ringkasan dengan konten informatif mencapai 82,48% untuk ringkasan 50% dan 93,76% untuk ringkasan 75% dari dokumen asli berdasarkan evaluasi Q&A. Dalam penelitian ini, hubungan antar kalimat ditetapkan dengan mengukur nilai kesamaan (*similarity*) menggunakan *content overlap.* Metrik tersebut mengandalkan perhitungan jumlah kata yang sama di antara kalimat-kalimat yang dibandingkan, tanpa memberikan bobot pada kata-kata tertentu. Dengan demikian, setiap kata diperlakukan setara dan dihitung secara sama dalam menentukan kesamaan antar kalimat. Penelitian yang dilakukan

oleh (Barrios *et al*., n.d.) yang berjudul "*Variations of the Similarity Function of TextRank for Automated Summarization*" mengkaji kekurangan *TextRank* dalam peringkasan teks otomatis. Kekurangan utama terletak pada fungsi kesamaan (*similarity function*) yang digunakan untuk membangun graf. Pada *TextRank*, kesamaan antara dua kalimat diukur berdasarkan banyaknya kata yang sama. Untuk mengatasi hal tersebut, penelitian ini mengajukan beberapa variasi fungsi kesamaan berbeda, seperti *Longest Common Substring*, BM25, dll. Dan Fungsi-fungsi kesamaan yang diusulkan mampu menangkap hubungan makna dan konteks yang lebih baik di antara kalimat, tidak hanya berdasarkan kesamaan kata. Dalam penelitian tersebu, beberapa variasi seperti BM25 dan BM25+ menunjukkan peningkatan signifikan mencapai 2,92% dalam skor ROUGE untuk evaluasi ringkasan dibandingkan metode *TextRank.* Penelitian yang dilakukan oleh (Cahyani & Patasik, 2021) yaitu perbandingan klasifikasi berita online menggunakan metode pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Term Frequency Absolute* (TF-ABS) dengan *algoritma Support Vector Machine* (SVM). Setelah melakukan penelitian dengan dataset berita *online* BBC sebanyak 2225 artikel, diperoleh hasil bahwa pembobotan TF-IDF memberikan kinerja yang lebih unggul dibandingkan TF-ABS ketika digabungkan dengan SVM. Metode TF-IDF dengan SVM mencapai rata-rata akurasi 96,63% dan f*1-score* 97,06%. Sedangkan TF-ABS dengan SVM hanya mendapat rata-rata akurasi 89,66% dan *f1-score* 96,63%. Terdapat kenaikan akurasi sebesar 6,97% dengan menggunakan TF-IDF dibanding TF-ABS. Dapat disimpulkan bahwa pembobotan TF-IDF lebih cocok dan efektif digunakan dalam penelitian ini.

Dari penelitian sebelumnya, penulis tertarik untuk mengembangkan *text summarization* yaitu dengan menggabungkan algoritma *TextRank* dengan pembobotan TF-IDF dalam artikel berbahasa indonesia dengan menerapkannya dalam aplikasi berbasis Web. Berdasarkan uraian diatas , penulis ingin melakukan penelitian berjudul "Implementasi *Text summarization* Pada Artikel Berbahasa Indonesia menggunakan Algoritma *TextRank* dengan *Term Frequency Inverse Document Frequency (*TF-IDF*)* berbasis Web". Diharapkan penelitian ini dapat membantu pelajar maupun peneliti untuk mendapatkan hasil ringkasan artikel yang baik dan akurat.

1.2 Identifikasi Masalah

Berdasarkan penjelasan latar belakang yang telah dibuat, maka permasalahan dalam penelitian ini dapat diidentifikasi sebagai berikut:

1. Pembaca memiliki waktu terbatas untuk memahami isi artikel, sehingga membutuhkan ringkasan artikel yang cepat dan efisien.
2. Aplikasi peringkasan teks yang tersedia saat ini memiliki kelemahan dalam mempertahankan integritas informasi penting dari teks sumber karena membuat kalimat-kalimat baru yang berbeda dari teks sumber asli.
3. Dibutuhkan sistem peringkasan teks yang dapat mempertahankan informasi kalimat penting dari teks sumber.

1.3 Ruang Lingkup

Berdasarkan penjelasan latar belakang yang telah dibuat, maka permasalahan dalam penelitian ini dapat diidentifikasi sebagai berikut:

1. Algoritma yang digunakan yaitu algoritma *TextRank* dengan TF-IDF.
2. Jenis *summarization* yang digunakan adalah *extractive summarization.*

1.4 Batasan Masalah

Batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Data yang digunakan dalam penelitian ini berbahasa Indonesia.
2. Artikel asli yang berformat PDF akan dikonversi terlebih dahulu menjadi dokumen teks (*plain text document*) berekstensi txt.
3. Abstrak dalam artikel dihapus.
4. Artikel yang digunakan mulai tahun 2022-2024 berupa file berekstensi pdf dari Jurnal Kode Unimed sebanyak 100 Jurnal.

1.5 Rumusan Masalah

Berdasarkan latar belakang di atas, maka dapat dirumuskan permasalahannya sebagai berikut:

1. Bagaimana implementasi *Text summarization* pada artikel berbahasa Indonesia menggunakan algoritma *TextRank dengan* TF-IDF berbasis Web?
2. Bagaimana hasil evaluasi *Text summarization* dengan menggunakan Algoritma *TextRank* dengan TF-IDF?

1.6 Tujuan Masalah

Tujuan Penelitian ini adalah sebagai berikut:

1. Mengetahui hasil implementasi *Text summarization* pada artikel berbahasa Indonesia menggunakan algoritma *TextRank* dengan TF-IDF berbasis Web.
2. Mengetahui hasil evaluasi *Text summarization* dengan menggunakan Algoritma *TextRank* dan TF-IDF.

1.7 Manfaat Penelitian

Penelitian ini dapat memberi manfaat antara lain:

1. Bagi peneliti, mendapat wawasan baru terkait *text summarization* dan membantu peneliti mempercepat proses pemahaman isi artikel dengan menghasilkan ringkasan secara otomatis dengan membuat aplikasi berbasis Web.
2. Bagi Pembaca, memudahkan untuk meringkas artikel berbahasa Indonesia secara cepat dan mudah diakses melalui aplikasi berbasis Web.
3. Bagi Universitas, penelitian ini dapat memberikan manfaat dalam dalam bahan studi bagi mahasiswa untuk penelitian kedepannya.