BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan cepat di bidang informatika dan ilmu komputer membuat jumlah artikel ilmiah yang dibuat meningkat sangat pesat. Fenomena ini terjadi di Indonesia maupun di berbagai belahan dunia karena permintaan akan inovasi teknologi semakin meningkat. Jumlah artikel jurnal yang dipublikasikan dalam bidang *Computer Science* telah meningkat signifikan dalam dua puluh tahun terakhir, dari sekitar 20.000 artikel pada tahun 2000 menjadi 120.000 artikel pada tahun 2023 (Smyth, 2025). Peningkatan besar ini, yang hampir mencapai enam kali lipat, didorong oleh tekanan akademik untuk *publish-or-perish* serta kemudahan akses melalui model *Open Access*, sehingga menciptakan *big data* literatur yang menantang bagi para peneliti dan akademisi.

Meskipun peningkatan jumlah publikasi menunjukkan kemajuan dalam ilmu pengetahuan, hal ini juga menimbulkan masalah baru, terutama pada tahap awal penelitian. Peneliti pemula atau mahasiswa seringkali harus menyortir ratusan abstrak untuk menemukan artikel yang relevan guna menyusun tinjauan pustaka (Liu et al., 2025). Dalam situasi ini, abstrak berfungsi sebagai dasar untuk menentukan relevansi penelitian. Namun, proses penyaringan literatur menjadi tidak efisien dan memakan waktu jika abstrak tersebut tidak dapat dibaca dengan baik, misalnya karena terlalu padat, mengandung jargon berlebihan, atau memiliki struktur kalimat yang kompleks. Transfer pengetahuan dan proses penelitian terhambat secara langsung oleh kesulitan menyaring literatur ini. Tingkat keterbacaan abstrak dalam jurnal ilmiah sangat berbeda, dan hal ini berdampak pada seberapa cepat dan mudah pembaca memahami isi karya (Jamar, 2025). Maka itu, dibutuhkan sistem otomatis yang mampu mengklasifikasikan tingkat kesulitan membaca abstrak agar peneliti dapat memprioritaskan atau menyaring literatur berdasarkan tingkat kemudahan pemahamannya.

Keterbacaan atau *readability* merupakan tingkat kemudahan suatu teks untuk dipahami oleh pembaca berdasarkan karakteristik linguistik yang dimilikinya. Tingkat keterbacaan dipengaruhi oleh struktur kalimat, panjang kata, kompleksitas

kosakata, serta susunan sintaksis yang membentuk suatu teks (Vajjala et al., 2020). Semakin tinggi kompleksitas unsur-unsur tersebut, semakin besar pula tingkat kesulitan pembaca dalam memahami isi teks. Dengan demikian, keterbacaan bukan hanya berkaitan dengan panjang teks, tetapi juga dengan bagaimana informasi disusun dan disampaikan secara linguistik. Dalam bidang Informatika yang mengalami peningkatan publikasi secara cepat, evaluasi keterbacaan abstrak menjadi sangat penting. Abstrak berperan sebagai ringkasan singkat yang berfungsi sebagai pintu masuk bagi para peneliti untuk memilih dan mengakses literatur. (Vajjala et al., 2020) menegaskan bahwa penilaian keterbacaan secara otomatis merupakan alat penting untuk mengatasi kelebihan informasi. Secara struktur, tingkat keterbacaan suatu dokumen dibentuk dari dua aspek utama yang dapat diukur secara numerik, yaitu kompleksitas leksikal dan kompleksitas sintaksis (Albrecht et al., 2021). Kompleksitas leksikal berkaitan dengan pemilihan dan panjang kata, sedangkan kompleksitas sintaksis mengacu pada struktur dan panjang kalimat. Dengan mengubah parameter leksikal dan sintaksis menjadi angka, sistem *machine learning* dapat dilatih untuk mengevaluasi tingkat kesulitan teks secara objektif dan mengurangi bias penilaian manual (Antić, 2021).

Untuk mengukur keterbacaan secara kuantitatif, beberapa metrik standar yang sering digunakan antara lain *Flesch-Kincaid Grade Level* (FKGL), *Flesch Reading Ease* (FRE), dan *Gunning Fog Index* (Gordejeva et al., 2022). Selain itu, *Automated Readability Index* (ARI) juga merupakan salah satu metode yang sering dipakai dalam analisis teks ilmiah dan teknis (Lee, 2025). ARI dikembangkan oleh Smith dan Senter pada tahun 1967 sebagaimana disebutkan dalam (Sarkar et al., 2020), dan memiliki keunggulan karena dirancang untuk dihitung secara otomatis sehingga cocok untuk analisis teks dalam skala besar. Berbeda dengan metode berbasis suku kata, ARI menggunakan rasio jumlah karakter per kata, sehingga lebih stabil dalam menganalisis teks teknis yang banyak mengandung akronim atau istilah panjang (Alamleh et al., 2025; Hamoda, 2025). Selain itu, ARI memberikan estimasi tingkat pendidikan pembaca yang selaras dengan konteks akademis (Fitzpatrick, 2023). Meskipun demikian, pendekatan berbasis formula seperti ARI pada dasarnya hanya menghasilkan skor numerik berdasarkan rumus tertentu dan belum melakukan pembelajaran pola dari data secara langsung. Skor yang

dihasilkan berupa angka mutlak dan belum tentu mampu menangkap variasi gaya bahasa teks yang beragam. Oleh karena itu, dibutuhkan pendekatan yang menggunakan data untuk menciptakan kategori tingkat kesulitan membaca secara otomatis berdasarkan bagaimana fitur-fitur dalam dataset tersebar.

Penelitian ini menawarkan metode menggunakan *machine learning* dengan memanfaatkan penggambaran kata dalam teks secara terperinci. Abstrak dokumen diubah menjadi matriks angka berisi fitur menggunakan metode TF-IDF dengan konfigurasi *unigram* dan *bigram*, sehingga dapat menangkap ciri teknis serta konteks bahasa yang digunakan. Fitur tersebut selanjutnya dibagi ke dalam kelompok menggunakan algoritma *K-Means Clustering* sehingga dapat membentuk kategori tingkat kesulitan membaca secara otomatis, yaitu mudah, sedang, dan sulit. Penggunaan *K-Means* dipilih karena metode ini efektif dalam mengelompokkan data berdasarkan kemiripan fitur numerik tanpa memerlukan label awal (*unsupervised learning*), serta mampu membentuk klaster dokumen teks yang solid setelah melalui tahap preprocessing yang tepat (Kurniawan, 2024). Selain itu, integrasi *K-Means* sebagai tahap pelabelan awal terbukti dapat meningkatkan kualitas proses klasifikasi pada tahap berikutnya (Sinatrya, 2025). Dengan demikian, kategori keterbacaan dibentuk secara objektif berdasarkan pola distribusi data, bukan berdasarkan batas ambang yang ditentukan secara subjektif. Selanjutnya, hasil pengelompokan digunakan sebagai label untuk membangun model klasifikasi menggunakan *Multinomial Naive Bayes* (MNB). Kemampuan MNB untuk menangani data teks yang diwakili dalam bentuk frekuensi kata atau fitur berbasis TF-IDF menentukan pemilihan MNB. MNB memiliki performa yang kompetitif dan efisien dalam klasifikasi teks berdimensi tinggi (Rachmatullah et al., 2025), serta terbukti mampu mencapai tingkat akurasi yang tinggi dalam berbagai penelitian klasifikasi teks (Mustofa et al., 2025). Karakteristik probabilistik yang sederhana namun efektif menjadikan MNB sesuai untuk mempelajari pola distribusi kata yang berkontribusi terhadap tingkat kesulitan suatu abstrak ilmiah.

Dalam penelitian ini, skor ARI tidak diposisikan sebagai satu-satunya penentu kategori keterbacaan, melainkan digunakan sebagai pendekatan pembanding atau baseline kuantitatif untuk mengevaluasi konsistensi hasil pengelompokan yang dihasilkan oleh metode berbasis *Clustering*. Oleh karena itu, penelitian ini tidak

hanya mengukur keterbacaan dengan menggunakan satu formula, tetapi juga membandingkan metode berbasis indeks keterbacaan dengan metode *machine learning* untuk menentukan sistem klasifikasi yang lebih objektif dan fleksibel..

1.2 Rumusan Masalah

"Merujuk pada pemaparan latar belakang di atas, rumusan masalah yang menjadi fokus dalam penelitian ini adalah sebagai berikut:

1. Bagaimana menggunakan algoritma *K-Means* yang didasarkan pada pembobotan fitur TF-IDF untuk secara objektif membentuk label tingkat keterbacaan (Mudah, Sedang, Sulit) pada abstrak ilmiah?
2. Bagaimana metode *Synthetic Minority Over-sampling Technique* (SMOTE) digunakan untuk membangun model klasifikasi *Multinomial Naive Bayes* (MNB) yang mampu menangani ketidakseimbangan jumlah data antar kelas?
3. Bagaimana performa model yang dihasilkan, dan sejauh mana tingkat keselarasan prediksinya jika dikomparasikan dengan standar *Automated Readability Index* (ARI)?

1.3 Batasan Masalah

Pembatasan masalah berikut dibuat dari rumusan masalah yang telah ditetapkan untuk membuat penelitian ini lebih fokus dan untuk mengukur hasilnya secara khusus:

1. Objek penelitian hanya fokus pada 3.000 abstrak jurnal ilmiah berbahasa Indonesia pada bidang Informatika yang bersumber dari portal Garba Rujukan Digital (GARUDA).
2. Analisis dilakukan hanya pada bagian abstrak dan tidak mencakup isi lengkap karya ilmiah.
3. Tahapan pra-pemrosesan teks dibatasi pada proses *case folding*, *cleansing*, dan *filtering*, serta tidak melibatkan tahapan *stemming* guna mempertahankan bentuk asli dari istilah-istilah teknis Informatika.
4. Metode *machine learning* yang digunakan dibatasi pada algoritma *K-Means Clustering* untuk pelabelan kategori otomatis dan *Multinomial Naive Bayes*

untuk klasifikasi, sedangkan skor *Automated Readability Index (ARI)* digunakan sebagai metode pembanding kuantitatif.

5. Hasil kategori keterbacaan mencerminkan tingkat kompleksitas bahasa teks dan tidak mempertimbangkan perbedaan tingkat pemahaman masing-masing pembaca maupun generalisasi ke bidang keilmuan lain.

1.4 Tujuan Penelitian

Mengacu pada perumusan masalah sebelumnya, capaian yang diharapkan dapat terealisasi dari penelitian ini dijabarkan sebagai berikut :

1. Menggunakan algoritma *K-Means* yang didasarkan pada pembobotan fitur TF-IDF, untuk membentuk label tingkat keterbacaan (Mudah, Sedang, Sulit) pada abstrak ilmiah secara objektif.
2. Membangun model klasifikasi Multinomial Naive Bayes (MNB) yang mampu menangani ketidakseimbangan jumlah data antar kelas dengan menggunakan teknik *Synthetic Minority Over-sampling Technique* (SMOTE).
3. mengevaluasi kinerja model klasifikasi yang dihasilkan dan tingkat keselarasan prediksinya dengan membandingkannya langsung dengan *Automated Readability Index* (ARI).

1.5 Manfaat Penelitian

Melalui penelitian ini, terdapat sejumlah manfaat teoretis maupun praktis yang diharapkan dapat dicapai, yaitu:

1. Bagi Peneliti (Penulis): Meningkatkan pemahaman dan kemampuan praktis di bidang *Machine Learning*, terutama dalam menerapkan algoritma *K-Means Clustering* dan *Multinomial Naive Bayes* untuk mengklasifikasikan teks.
2. Bagi Pengguna (Mahasiswa dan Akademisi): Memberikan sistem klasifikasi teks secara otomatis yang bersifat objektif, sehingga membantu mahasiswa dan peneliti dalam melakukan tinjauan pustaka, mempercepat proses penyaringan literatur, dan meningkatkan efisiensi dalam melakukan penelitian.

1.6 Metode Penelitian

Penelitian ini fokus pada pengembangan model klasifikasi yang melibatkan beberapa tahapan berikut :

1. Pengumpulan Data

   Melakukan ekstraksi data sekunder berupa 3.000 abstrak jurnal ilmiah Informatika berbahasa Indonesia dari portal GARUDA melalui teknik *web scraping* menggunakan Selenium.

2. Pra-Pemrosesan Data

   Membersihkan dan menyeragamkan data teks melalui tahapan *case folding*, *cleansing* untuk menghapus karakter selain huruf, serta *filtering* untuk membuang kata hubung bahasa Indonesia dan Inggris, sehingga data siap diolah.

3. Ekstraksi Fitur

   Mengubah data teks hasil pra-pemrosesan menjadi bentuk angka melalui pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF) menggunakan rentang *unigram-bigram*, serta menyeleksi 3.000 fitur tertinggi agar karakteristik leksikal teks terpetakan dengan jelas.

4. Pelabelan Otomatis (*Auto-labeling*)

   Menerapkan algoritma *K-Means Clustering* ($k = 3$) pada hasil reduksi dimensi fitur TF-IDF menggunakan TruncatedSVD untuk membentuk tiga kelompok dokumen. Selanjutnya, setiap klaster diberi label Mudah, Sedang, dan Sulit berdasarkan nilai rata-rata jumlah kata per kalimat.

6. Pembagian dan Penyeimbangan Data dengan SMOTE

   Mengatasi potensi ketidakseimbangan jumlah sampel pada kategori minoritas (data latih) menggunakan teknik SMOTE (*Synthetic Minority Over-sampling Technique*) dengan rasio 80:20 (latih:uji) guna memastikan model tidak bias.

7. Klasifikasi *Multinomial Naïve Bayes*

   Membangun model klasifikasi menggunakan algoritma MNB (*alpha* = 0.1) dengan masukan fitur TF-IDF dan label hasil pengelompokan otomatis untuk menghasilkan fungsi prediksi yang adaptif terhadap data abstrak baru.

8. Evaluasi dan Validasi

   Mengevaluasi kinerja komputasional model melalui pengukuran akurasi, presisi, *recall*, dan *F1-score*. Selain itu, melakukan validasi silang eksternal dengan menggunakan perhitungan *Automated Readability Index* (ARI) sebagai dasar untuk mengevaluasi tingkat keterbacaan.

1.7 Sistematika Penulisan

Sistematika penulisan penelitian ini disusun secara sistematis sehingga pembaca lebih mudah memahami alur pembahasan. Sistematika penulisan penelitian ini terdiri dari beberapa bab berikut :

BAB I PENDAHULUAN

Bab ini menjelaskan latar belakang, rumusan masalah, batasan, tujuan penelitian, manfaat penelitian, dan gambaran umum metode penelitian sebagai arah dasar penelitian ini.

BAB II TINJAUAN PUSTAKA

Bab ini menguraikan landasan teori yang mendukung penelitian, meliputi konsep keterbacaan teks, ekstraksi fitur TF-IDF, algoritma *K-Means Clustering*, *Truncated Singular Value Decomposition* (TruncatedSVD), SMOTE, *Multinomial Naive Bayes* (MNB), *Automated Readability Index* (ARI), serta pembahasan mengenai riset-riset pendahulu yang mendukung penelitian ini.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan langkah-langkah teknis yang digunakan dalam penelitian. Ini mencakup pengumpulan data, pra-pemrosesan teks, ekstraksi fitur TF-IDF, pengurangan dimensi dengan TruncatedSVD, pelabelan otomatis menggunakan K-*K-Means Clustering*, pembagian data, penyeimbangan data dengan SMOTE, klasifikasi menggunakan *Multinomial Naive Bayes*, dan proses evaluasi dan validasi hasil klasifikasi.

BAB IV HASIL DAN PEMBAHASAN

Bab ini memaparkan rincian temuan serta analisis dari keseluruhan prosedur penelitian. Pembahasan di dalamnya mencakup proses akuisisi dan pembersihan data, konversi fitur *TF-IDF*, pelabelan otomatis dengan *K-Means Clustering*,

penyetaraan sampel menggunakan SMOTE, implementasi model *Multinomial Naive Bayes*, penilaian kinerja sistem, dan diakhiri dengan verifikasi komparatif merujuk pada *Automated Readability Index* (ARI).

BAB V PENUTUP

Bab ini menjadi penutup yang menyajikan jawaban akhir atas rumusan masalah penelitian, ditambah dengan rekomendasi operasional bagi peneliti yang ingin mengembangkan topik ini lebih lanjut.