

BAB I PENDAHULUAN

1.1 Latar Belakang

Perkembangan internet yang begitu pesat membawa perubahan dalam kehidupan, salah satunya dalam penyebaran informasi. Perubahan ini ditandai dengan munculnya berbagai platform media social yang memungkinkan informasi menyebar secara instan dimanapun dan kapanpun. Transformasi ini telah menjadikan masyarakat semakin bergantung pada media daring sebagai sumber utama berita dan pengetahuan. Namun, kemudahan berbagi ini juga menciptakan tantangan baru yang signifikan, yaitu maraknya penyebaran informasi palsu atau berita hoaks [1]. Dampak dari hoaks ini tidak hanya berdampak pada individu saja melainkan dapat memecah belah masyarakat, dan bahkan mengancam proses demokrasi. Urgensi masalah ini menuntut adanya solusi teknologi yang proaktif, bukan hanya reaktif, dalam menangani volume berita yang terus berlipat ganda.

Menghadapi tantangan ini, pendekatan berbasis Natural Language Processing dan Machine Learning menawarkan mekanisme yang terukur dan otomatis untuk klasifikasi berita. Tugas mendeteksi hoaks dapat diformulasikan sebagai masalah klasifikasi biner, di mana sistem dilatih untuk membedakan secara otomatis antara berita Hoaks dan bukan Hoaks berdasarkan konten tekstualnya. Keberhasilan proses ini sangat bergantung pada bagaimana teks mentah dapat direpresentasikan menjadi data numerik yang bermakna. *Term Frequency-Inverse Document Frequency (TF-IDF)* muncul sebagai teknik ekstraksi fitur yang efektif dan *robust* [2]. TF-IDF menghitung nilai bobot untuk setiap kata, semakin spesifik suatu kata muncul dalam sebuah dokumen frekuensi tinggi tetapi jarang muncul di seluruh korpus dokumen frekuensi rendah, semakin tinggi bobotnya. Representasi ini sangat vital karena membantu algoritma untuk fokus pada kata-kata kunci unik yang sering menjadi penanda pola berita palsu dalam data berbahasa Indonesia.

Dalam algoritma klasifikasi, Naive Bayes dan Logistic Regression menonjol karena keseimbangan antara efisiensi komputasi dan kinerja yang andal pada tugas klasifikasi teks berdimensi tinggi. Naive Bayes, sebagai model generatif yang didasarkan pada Teorema Bayes, populer karena kesederhanaannya dan kecepatan pelatihannya [3]. Sebaliknya, Logistic Regression adalah model diskriminatif linier yang memodelkan probabilitas kelas secara langsung dan seringkali dapat mengatasi keterbatasan asumsi independensi fitur yang dimiliki

oleh NB, sehingga berpotensi memberikan akurasi yang lebih baik dalam kondisi tertentu[4].

Banyak penelitian terdahulu telah membandingkan kedua model ini pada berbagai klasifikasi teks, dengan hasil yang bervariasi. Misalnya, beberapa studi menunjukkan bahwa Logistic Regression lebih unggul [4], sementara studi lain menyimpulkan bahwa Naive Bayes tetap menjadi *baseline* yang efektif dengan waktu komputasi yang lebih singkat [5]. Namun, kinerja model sangat sensitif terhadap bahasa, kualitas *preprocessing*, dan teknik ekstraksi fitur yang digunakan. Oleh karena itu, penelitian ini bertujuan untuk memberikan analisis komparatif yang terfokus, secara eksplisit membandingkan Model Naive Bayes dan Logistic Regression hanya dengan menggunakan fitur TF-IDF sebagai *baseline* fitur yang kuat dan sederhana pada korpus berita hoaks Bahasa Indonesia. Kebaruan penelitian ini terletak pada upaya sistematis untuk mengidentifikasi model mana (NB atau LR) yang menunjukkan kinerja paling seimbang dilihat dari *Accuracy*, *Precision*, *Recall*, dan *F1-score*, sebagai dasar kuat untuk implementasi sistem deteksi hoaks yang efisien di Indonesia.

1.2 Rumusan Masalah

Berdasarkan uraian latar belakang dan identifikasi masalah yang telah disampaikan, maka dapat dirumuskan permasalahan penelitian ini sebagai berikut:

1. Bagaimana kinerja Model Naive Bayes dan Logistic Regression dalam mendeteksi berita hoaks berbahasa Indonesia yang direpresentasikan menggunakan fitur TF-IDF?
2. Model manakah antara Naive Bayes atau Logistic Regression yang memberikan performa paling optimal untuk deteksi berita hoaks berbahasa Indonesia, berdasarkan metrik evaluasi seperti *Accuracy*, *Precision*, *Recall*, dan *F1-Score*?

1.3 Batasan Masalah

Mengingat luasnya ruang lingkup permasalahan yang berkaitan dengan deteksi berita hoaks, penelitian ini perlu difokuskan pada aspek-aspek tertentu agar hasil yang diperoleh lebih akurat dan mendalam. Oleh karena itu, batasan masalah dalam penelitian ini ditetapkan sebagai berikut:

1. Dataset yang digunakan terbatas pada data berita bahasa Indonesia yang sudah diklasifikasikan kedalam dua kategori yaitu hoaks dan bukan hoaks.
2. Penelitian ini dibatasi pada penggunaan dataset berita yang bersumber dari portal berita daring CNN Indonesia, Detik.com, dan Kompas.com sebagai representasi berita valid (non-hoaks), serta TurnBackHoax.id sebagai sumber dataset berita hoaks, sehingga data

dari sumber lain di luar portal tersebut tidak termasuk dalam penelitian ini.

3. Metode yang digunakan untuk merepresentasikan dan membobotkan fitur teks (kata) dari korpus data adalah *Term Frequency-Inverse Document Frequency* (TF-IDF). Metode ekstraksi fitur lain tidak digunakan.
4. Model Machine Learning yang dianalisis dan dibandingkan performanya hanya dua, yaitu model Naive Bayes dan Logistic Regression.

1.4 Tujuan Penelitian

Sejalan dengan rumusan masalah yang telah ditetapkan, penelitian ini memiliki tujuan untuk mencapai hasil yang dapat menjawab permasalahan tersebut, yaitu:

1. Untuk menganalisis dan menghitung kinerja model Naive Bayes dan Logistic Regression dalam mendeteksi berita hoaks berbahasa Indonesia dengan fitur TF-IDF untuk pembobotan kata.
2. Untuk membandingkan performa kedua model dan mengidentifikasi model yang paling optimal untuk tugas deteksi berita hoaks Bahasa Indonesia, berdasarkan hasil evaluasi *Accuracy*, *Precision*, *Recall*, dan *F1-Score*.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat baik secara teoritis maupun praktis, yang dijelaskan sebagai berikut:

1. Manfaat Teoritis
 - a. Penelitian ini diharapkan dapat memperkaya literatur dan referensi ilmiah, khususnya dalam bidang *Natural Language Processing* (NLP) dan Machine Learning, dengan fokus pada isu berita hoaks dalam konteks Bahasa Indonesia.
 - b. Hasil analisis perbandingan yang terstruktur antara model Naive Bayes dan Logistic Regression berbasis TF-IDF dapat memberikan temuan empiris yang kuat mengenai karakteristik dan kinerja kedua model pada data teks Bahasa Indonesia.
2. Manfaat Praktis
 - a. Hasil penelitian ini memberikan rekomendasi yang spesifik dan terukur mengenai model klasifikasi Naive Bayes atau Logistic Regression yang paling optimal untuk diimplementasikan dalam pengembangan sistem atau aplikasi *filtering* berita hoaks di Indonesia.
 - b. Kontribusi penelitian dapat dimanfaatkan oleh instansi atau platform yang bertanggung jawab atas pengawasan konten misalnya Kominfo, atau penyedia

platform media sosial sebagai masukan data untuk meningkatkan akurasi sistem deteksi otomatis mereka.

- c. Secara tidak langsung, dengan membantu mengidentifikasi model yang efisien, penelitian ini mendukung upaya yang lebih luas dalam meningkatkan kualitas informasi daring dan memfasilitasi lingkungan digital yang lebih sehat bagi masyarakat Indonesia.

1.6 Sistematika Penulisan

Penelitian ini disusun dalam beberapa bab agar pembahasan tersusun secara sistematis dan mudah dipahami. Adapun sistematika penulisan dalam penelitian ini adalah sebagai berikut :

BAB I PENDAHULUAN, Bab ini membahas latar belakang penelitian, perumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, serta sistematika penulisan sebagai gambaran umum alur penelitian.

BAB II TINJAUAN PUSTAKA, Bab ini berisi landasan teori yang mendukung penelitian, meliputi konsep berita hoaks, *text mining*, praproses teks berbahasa Indonesia, TF-IDF, algoritma Naïve Bayes, dan Logistic Regression. Selain itu, bab ini juga memuat penelitian terdahulu yang relevan serta perbandingan penelitian untuk menunjukkan posisi dan kebaruan penelitian yang dilakukan.

BAB III METODE PENELITIAN, Bab ini menjelaskan tahapan penelitian secara rinci, mulai dari pengumpulan dataset berita hoaks berbahasa Indonesia, proses praproses teks, ekstraksi fitur menggunakan TF-IDF, pemodelan klasifikasi menggunakan Naïve Bayes dan Logistic Regression, skenario pembagian data latih dan data uji, hingga metode evaluasi kinerja model.

BAB IV HASIL DAN PEMBAHASAN, Bab ini menyajikan hasil pengujian model Naïve Bayes dan Logistic Regression pada setiap skenario pengujian yang dilakukan. Selanjutnya, dilakukan analisis dan pembahasan terhadap hasil evaluasi yang diperoleh untuk menilai performa, stabilitas, dan konsistensi masing-masing model dalam mendeteksi berita hoaks.

BAB V PENUTUP, Bab ini berisi kesimpulan yang diperoleh berdasarkan hasil penelitian serta saran untuk pengembangan penelitian selanjutnya, khususnya terkait penggunaan dataset yang lebih besar, penerapan metode lain, atau pengembangan sistem deteksi berita hoaks yang lebih akura.