202010370311315
Doni Ramadhani
Prodi Informatika

BAB I
PENDAHULUAN

1.1. Latar Belakang

Perkembangan teknologi kecerdasan buatan (Artificial Intelligence/AI) telah memberikan dampak yang signifikan dalam berbagai aspek di kehidupan manusia. AI berkembang pesat dan telah diterapkan secara luas di berbagai sektor. Di bidang kesehatan, misalnya, AI dimanfaatkan untuk menganalisis citra medis dengan tingkat akurasi tinggi. Di sektor keuangan, teknologi ini digunakan untuk deteksi penipuan secara real-time serta membantu dalam pengelolaan risiko keuangan. Sementara itu, di sektor industri dan manufaktur, AI berperan dalam mengotomatisasi proses produksi, serta meningkatkan efisiensi dan produktivitas[1]. AI memungkinkan sistem komputer untuk menjalankan tugas-tugas kompleks yang sebelumnya hanya dapat dilakukan oleh manusia, seperti pengenalan suara, pengambilan keputusan, serta pemrosesan dan pemahaman bahasa alami.

Natural Language Processing (NLP) adalah cabang dari kecerdasan buatan (Artificial Intelligence/AI) yang berfokus pada interaksi antara manusia dan komputer menggunakan bahasa alami. NLP memungkinkan komputer untuk memahami, memproses, dan menghasilkan teks atau ucapan manusia secara otomatis dan bermakna. Dengan berkembangnya teknologi, NLP telah menjadi komponen penting dalam berbagai aplikasi modern seperti chatbot, penalaran otomatis, analisis sentimen, klasifikasi dokumen, serta pencarian informasi berbasis teks.[1][2][3][4].

Salah satu penerapan NLP yang umum digunakan adalah dalam klasifikasi kategori berita, yaitu proses pengelompokan teks berita ke dalam topik-topik tertentu seperti politik, ekonomi, olahraga, hiburan, dan lain sebagainya[3][4]. Penerapan ini sangat penting untuk menyaring informasi secara otomatis, mengatur konten berita, serta meningkatkan efisiensi dalam pencarian dan rekomendasi berita. Teks berita merupakan jenis data yang sangat cocok untuk klasifikasi terutama untuk kasus pengujian stemming, karena umumnya ditulis dalam bahasa yang baku dan terstruktur, sehingga memudahkan proses ekstraksi fitur oleh sistem klasifikasi.

Klasifikasi teks merupakan proses pengelompokan teks ke dalam kategori atau kelas tertentu berdasarkan isi dan makna yang dikandungnya. Proses ini umumnya dilakukan dengan memanfaatkan teknik pembelajaran mesin (Machine Learning) untuk menganalisis teks dan

menentukan label yang paling relevan. Berbagai algoritma klasifikasi telah dikembangkan, masing-masing dengan karakteristik dan mekanisme kerja yang berbeda-beda. Dalam klasifikasi teks, terdapat dua pendekatan utama, yaitu supervised learning dan unsupervised learning [5]. Pemilihan algoritma Naïve Bayes dalam penelitian ini juga didasarkan pada kesesuaiannya dengan fokus topik, yaitu efektivitas stemming pada klasifikasi teks. Beberapa penelitian sebelumnya telah menunjukkan bahwa Naïve Bayes sangat cocok dipadukan dengan metode stemming berbasis aturan seperti Enhanced Confix Stripping (ECS), khususnya dalam konteks data berita. Penelitian oleh Erwin Yudi Hidayat et al. (2020) menunjukkan bahwa kombinasi ECS dan Naïve Bayes mampu mencapai akurasi hingga 95% dalam klasifikasi dokumen berita[6]. Sementara itu, studi oleh Yoga Dwitya Pramudita et al. (2018) yang menerapkan kombinasi yang sama untuk klasifikasi berita olahraga juga memperoleh akurasi yang baik, yaitu sebesar 77%[7]. Kedua penelitian tersebut membuktikan bahwa Naïve Bayes merupakan algoritma yang relevan dan efektif untuk digunakan dalam studi evaluasi stemming terhadap teks berita berbahasa Indonesia.

Dalam klasifikasi teks, kualitas data memegang peranan penting dalam menentukan performa model. Oleh karena itu, tahapan preprocessing sangat diperlukan untuk memastikan bahwa data yang digunakan bersih, terstruktur, dan siap untuk diproses oleh algoritma klasifikasi. Tahapan preprocessing umumnya mencakup case folding, cleansing, stopword removal, tokenizing, dan stemming [8]. Salah satu tantangan utama dalam pemrosesan teks, khususnya dalam Bahasa Indonesia, adalah keberagaman bentuk kata yang disebabkan oleh penggunaan imbuhan, awalan, dan akhiran. Oleh karena itu, proses stemming menjadi salah satu langkah krusial dalam preprocessing untuk menyederhanakan bentuk kata ke dalam bentuk dasarnya.

Stemming merupakan proses mengubah kata menjadi bentuk dasarnya [9]. Stemming membantu mengurangi jumlah fitur unik dalam teks dengan mengelompokkan kata-kata yang memiliki makna serupa. Misalnya, kata *"berlari"*, *"lari"*, dan *"berlarian"* akan direduksi menjadi *"lari"*, sehingga model lebih efisien dalam mengenali pola dan meningkatkan akurasi klasifikasi.

Berbagai metode stemming telah dikembangkan untuk bahasa Indonesia, di antaranya adalah Enhanced Confix Stripping (ECS) dan IN-Idris [6][10]. ECS merupakan pengembangan dari algoritma Confix Stripping berbasis aturan (rule-based), yang bekerja dengan memotong akhiran terlebih dahulu, kemudian dilanjutkan dengan pemotongan atau perubahan awalan

apabila diperlukan. Metode Enhanced Confix Stripping (ECS) dilengkapi dengan mekanisme pengembalian akhiran untuk mengatasi permasalahan overstemming, yaitu kondisi ketika kata dasar yang dihasilkan menjadi terlalu sederhana sehingga kehilangan makna aslinya [11][6]. Berdasarkan penelitian sebelumnya, Enhanced Confix Stripping (ECS) menunjukkan kinerja yang sangat baik dalam memproses data berita [6].

Sementara itu, IN-Idris dipilih karena merupakan versi penyempurnaan dari Idris Stemmer sama seperti Enhanced confix stripping yang menyempurnakan confix stripping stemmer, memperbaiki aturan-aturan sebelumnya guna menghasilkan kata dasar dengan tingkat akurasi yang lebih tinggi. Algoritma IN-Idris memulai prosesnya dengan menghapus atau merubah awalan terlebih dahulu, kemudian dilanjutkan dengan penghapusan akhiran jika diperlukan. Jika kata dasar masih belum ditemukan, proses penghapusan imbuhan akan diulangi. Penelitian sebelumnya menunjukkan bahwa performa IN-Idris mampu melampaui algoritma Idris, dengan peningkatan akurasi sebesar 5,25% [10].

Meskipun keduanya mengadopsi pendekatan berbasis aturan (rule-based), Enhanced Confix Stripping (ECS) dan IN-Idris memiliki perbedaan dalam langkah pemotongan kata serta aturan yang diterapkan, seperti pemotong awalan yang dilakukan terlebih dahulu dan memiliki pengulangan algoritma pada IN-Idris sedangkan ECS melakukan pemotongan akhiran terlebih dahulu dan memiliki loop pengembalian akhiran[10][11]. Perbedaan ini menjadi aspek menarik untuk diteliti, mengingat keduanya memiliki tujuan yang sama, yaitu mengubah kata menjadi bentuk dasarnya, namun melalui prosedur dan aturan yang berbeda. Oleh karena itu, penelitian ini bertujuan untuk menganalisis perbedaan hasil stemming yang dihasilkan oleh kedua algoritma tersebut serta mengevaluasi dampaknya terhadap akurasi model klasifikasi teks. Selain itu, penelitian ini juga berfokus pada identifikasi kecenderungan masing-masing metode terhadap fenomena overstemming maupun understemming, yang dapat mempengaruhi kualitas hasil klasifikasi.

Hasil dari penelitian ini diharapkan dapat memberikan wawasan mengenai efektivitas masing-masing metode stemming dalam klasifikasi teks menggunakan algoritma Naïve Bayes, serta memberikan rekomendasi kepada peneliti dan praktisi dalam memilih teknik praproses (preprocessing) yang optimal untuk pemrosesan teks berbahasa Indonesia. Selain itu, temuan dari penelitian ini juga diharapkan dapat menjadi dasar bagi pengembangan metode stemming yang lebih adaptif terhadap kompleksitas struktur bahasa Indonesia di masa mendatang.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, penelitian ini akan berfokus pada analisis perbedaan dan efektivitas dua algoritma stemming, yaitu Enhanced Confix Stripping (ECS) dan IN-Idris, dalam pemrosesan teks berbahasa Indonesia. Adapun rumusan masalah dalam penelitian ini adalah sebagai berikut:

a. Bagaimana perbedaan hasil stemming antara algoritma Enhanced Confix Stripping dan IN-Idris dalam pemrosesan teks bahasa Indonesia?

b. Bagaimana pengaruh penggunaan kedua algoritma stemming tersebut terhadap performa model klasifikasi teks berbasis machine learning?

c. Sejauh mana algoritma Enhanced Confix Stripping dan IN-Idris mampu mengatasi kata berimbuhan dalam bahasa Indonesia, serta bagaimana efektivitasnya dalam mengurangi masalah overstemming dan understemming?

1.3. Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah sebagai berikut:

a. Menganalisis perbedaan hasil stemming yang dihasilkan oleh algoritma Enhanced Confix Stripping dan IN-Idris dalam pemrosesan teks bahasa Indonesia.

b. Menilai pengaruh penggunaan algoritma Enhanced Confix Stripping dan IN-Idris terhadap performa model klasifikasi teks berbasis machine learning.
c. Mengevaluasi efektivitas algoritma Enhanced Confix Stripping dan IN-Idris dalam menangani kata berimbuhan.

1.4. Batasan Masalah

Agar penelitian ini lebih terarah dan mendalam, beberapa batasan masalah yang ditetapkan adalah sebagai berikut:

a. Penelitian ini hanya menggunakan algoritma Naïve Bayes sebagai model klasifikasi teks.

b. Data yang digunakan dalam penelitian ini berasal dari data berita KOMPASTV yang besisi judul dan kategori berita.

c. Metode representasi fitur yang digunakan dalam klasifikasi teks terbatas pada eknik *Bag of Words* (BoW).

d. Evaluasi penelitian ini berfokus pada dua aspek utama, yaitu: (a) tingkat akurasi model klasifikasi teks yang dihasilkan setelah penerapan algoritma stemming, serta (b) efektivitas algoritma stemming dalam menangani kata berimbuhan, khususnya dalam mengatasi permasalahan *overstemming* dan *understemming*.

Dengan batasan ini, penelitian diharapkan dapat menghasilkan analisis yang lebih terarah serta memberikan rekomendasi yang aplikatif dalam pemilihan metode stemming untuk klasifikasi teks dalam bahasa Indonesia.

1.5. Sistematika Penulisan

Sistematika penulisan laporan penelitian ini disusun menjadi beberapa bab sebagai berikut:

BAB I PENDAHULUAN

Pada bab ini berisi pendahuluan yang menjelaskan latar belakang, perumusan masalah, tujuan, batasan masalah, dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Bab ini berisi mengenai kajian pustaka sebagai parameter rujukan untuk dilaksanakannya penelitian.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan tentang tahapan desain penelitian, kerangka, dan konsep penelitian yang digunakan untuk menyelesaikan permasalahan penelitian.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menjelaskan mengenai implementasi, pengujian, hasil penelitian serta pembahasan mengenai hasil penelitian. pengujian membuat implementasi meliputi implementasi algoritma dan hasil simulasi, hasil pengujian simulasi meliputi skenario perbandingan stemming.

BAB V PENUTUP

Bab ini berisikan kesimpulan dari sistem yang dibuat serta saran untuk kepentingan yang lebih lanjut.