

BAB I

PENDAHULUAN

1.1 Latar Belakang

Pemindahan Ibu Kota Negara (IKN) dari Jakarta ke Kalimantan Timur yang diberi nama “Nusantara”, pertama kali diumumkan oleh Presiden Joko Widodo pada 29 April 2019, merupakan bagian dari upaya untuk mengatasi ketimpangan pembangunan antara Pulau Jawa dan wilayah lain di Indonesia. Nusantara dipilih sebagai lokasi Ibu Kota Negara karena letaknya yang strategis[1]. Namun, perpindahan ini menimbulkan pro dan kontra di masyarakat. Beberapa pihak mendukung langkah ini sebagai upaya strategis untuk memacu pertumbuhan ekonomi di luar Jawa, sementara pihak lain mengkritik pemilihan lokasi dan pengesahan undang-undang yang dinilai terburu-buru. Selain itu, muncul kekhawatiran tentang biaya besar yang diperlukan untuk pembangunan IKN, serta dampaknya terhadap lingkungan dan masyarakat lokal di Kalimantan[2]. Meskipun demikian, pemerintah tetap optimis bahwa IKN akan menjadi simbol kemajuan Indonesia, dengan konsep yang modern dan ramah lingkungan, serta mampu mengurangi beban Jakarta sebagai pusat pemerintahan dan bisnis[1].

Topik tersebut menjadi sangat hangat diperbincangkan oleh kalangan masyarakat terutama di platform media sosial seperti Tiktok, Instagram dan Youtube[1], [2]. Tidak hanya itu, pemerintah juga menggunakan media sosial untuk meningkatkan pemahaman publik tentang manfaat relokasi IKN dan menggandeng *influencer* untuk mengkampanyekan proyek ini secara positif. Hal ini sejalan dengan penggunaan media sosial sebagai alat promosi digital yang efektif dalam memasarkan produk maupun ide[3]. Pengguna dari berbagai latar belakang bebas mengekspresikan dukungan maupun kritik terhadap kebijakan pemerintah ini, baik dalam bentuk komentar, unggahan video, hingga konten diskusi panjang[4]. Media sosial juga memfasilitasi penyebaran informasi dan mempercepat penyusunan opini publik tentang IKN, menciptakan ruang diskusi yang sering kali sangat dinamis[5]. Sehingga banyak opini baik positif maupun negatif dalam menyikapi kebijakan pemerintah ini[6]. Namun, kesalahan penulisan (*typo*) pada sentimen yang diungkapkan pada media sosial seperti Tiktok, Instagram maupun Youtube oleh

masyarakat kerap dan wajar terjadi. Kesalahan penulisan tersebut sering terjadi dikarenakan kelalaian pengguna sendiri serta keinginan pengguna untuk mempersingkat kata agar tidak terlalu panjang ketika diutarakan namun masih bisa terbaca oleh pengguna lain. Yang tentunya hal tersebut akan berpengaruh terhadap keakuratan klasifikasi. Untuk mengatasi potensi pengaruh kesalahan penulisan, perlu metode tambahan untuk melakukan koreksi ejaan terhadap kesalahan penulisan kata tersebut[7].

Damerau-Levenshtein Distance bisa digunakan untuk mengantisipasi permasalahan kesalahan penulisan yang diungkapkan tersebut dengan baik. Metode ini menentukan jumlah minimum operasi yang dibutuhkan untuk mengubah satu string menjadi string lain, di mana operasi yang digunakan yaitu *insertion*, *deletion*, *substitution* dan *transposition* diantara dua karakter[8]. Penggunaan metode ini sangat efektif dan bisa meningkatkan akurasi mencapai 9% dalam kalimat dan kata pada teks pengujian yang memiliki ejaan tidak sah[9], [10]. Penggunaan dari metode koreksi ejaan tersebut akan sangat berguna jika diterapkan pada *text mining*, yakni guna peningkatan akurasi model[7]. *Text Mining* adalah proses ekstraksi informasi dari data teks yang tidak terstruktur, yang memungkinkan analisis data dalam skala besar. *Natural Language Processing* (NLP) berfungsi untuk mengidentifikasi dan mengekstrak opini dari teks, sehingga dapat mengubah konten teks menjadi data kuantitatif yang berguna untuk pengambilan keputusan[11]. Analisis Sentimen merupakan salah satu aplikasi dari *text mining* yang bertujuan untuk mengklasifikasikan opini dalam teks menjadi kategori positif, negatif, atau netral. *Natural Language Preprocessing* adalah langkah awal yang penting dalam analisis sentimen, yang mencakup berbagai teknik untuk mempersiapkan data teks agar siap untuk dianalisis[12].

Algoritma *Random Forest Classifier* merupakan metode klasifikasi dalam bidang *machine learning* dan bisa digunakan untuk analisis sentimen, yang memiliki performa tinggi dibandingkan dengan metode klasifikasi lainnya serta memiliki kelemahan terkait akurasi dan stabilitas data yang relatif rendah. Ini mengacu pada fungsi acak yang dihasilkan untuk mengacak pemilihan baris data dan pemecah atribut kandidat[13]. Maka diperlukan ekstraksi fitur dan juga seleksi

fitur untuk meningkatkan akurasi dan meningkatkan stabilitas data dari algoritma tersebut.

Ekstraksi fitur yang biasa digunakan ialah *N-Gram* dan TF-IDF. TF-IDF akan menghitung bobot kata yang dapat digunakan untuk mengetahui pola dari sebuah opini[14]. *N-Gram* akan memecah kata menjadi unit-unit yang lebih kecil seperti *unigram*, *bigram* dan *trigram*[15]. Permasalahan yang sering terjadi dalam analisis sentimen adalah ketika model hanya mempertimbangkan kata-kata secara individual (*unigram*), sehingga bisa menghasilkan interpretasi yang salah. Sebagai contoh, dalam ulasan ‘suka hutang’, jika dipecah menjadi dua kata, ‘suka’ akan terdeteksi sebagai sentimen positif dan ‘hutang’ sebagai sentimen negatif. Namun, kedua kata ini sebenarnya membentuk konteks yang berbeda ketika digabungkan. Dengan menggunakan *N-Gram*, misalnya *Bigram*, frasa ‘suka hutang’ dapat dianalisis sebagai satu kesatuan, yang seharusnya terdeteksi sebagai sentimen negatif secara keseluruhan, sesuai dengan makna yang dimaksud dalam konteks kalimat tersebut. Kedua metode tersebut dapat dipakai secara bersamaan *N-Gram* digunakan untuk tokenisasi dilanjutkan dengan pembobotan kata menggunakan TF-IDF seperti pada penelitian yang dilakukan oleh Y. Hapsari et. al [16]. Selain itu, seleksi fitur *Information Gain* juga dapat berperan penting dalam meningkatkan kinerja algoritma *Random Forest*. Sebagai teknik seleksi fitur, *Information Gain* memilih fitur-fitur terbaik dengan mengekstrak kata-kata kunci dari hasil pengurangan fitur[17]. Dalam konteks analisis sentimen, metode kerangka kerja dalam *text mining* ataupun *data mining* seperti CRISP-DM dapat digunakan untuk memberikan pendekatan yang lebih terstruktur dan efektif. CRISP-DM sebagai kerangka kerja, menyediakan panduan sistematis dalam menangani kompleksitas analisis sentimen. Dengan menerapkan CRISP-DM, proses seperti pemahaman bisnis (*bussiness understanding*), pemahaman data(*data understanding*), pemodelan(*modelling*), evaluasi(*evaluation*), dan pengembangan(*deployment*) dapat dijalankan secara terencana serta sistematis untuk meningkatkan performa algoritma *Random Forest* sekaligus memastikan konsistensi dalam hasil yang dihasilkan[18].

Penelitian yang menggunakan metode klasifikasi dengan menggunakan koreksi ejaan sudah banyak digunakan. Penelitian yang dilakukan oleh D. Irawan

et. al [19] menggunakan algoritma *Random Forest* menghasilkan akurasi 96,8% lebih tinggi dibandingkan dengan metode *Naïve Bayes*, dengan metode perbaikan kata *Damerau-Levenshtein Distance*, pada penelitian [20] menghasilkan peningkatan akurasi klasifikasi, mencapai 0,9. Penelitian lainnya, penggunaan *Damerau-Levenshtein Distance* [8] dan *N-Gram* [21] menghasilkan saran akurasi yang baik dengan akurasi mencapai diatas 85%. Selain itu, penelitian [18] dengan penambahan metodologi kerangka kerja CRISP-DM diperoleh hasil akurasi sebesar 98,46% dan meningkatkan kinerja algoritmik dalam proses klasifikasi. Oleh karena itu, penelitian ini menganalisis sentimen mengenai relokasi Ibu Kota Nusantara. Sentimen masyarakat ini diambil dari komentar dan pendapat yang diungkapkan pada sosial media Tiktok, Youtube dan Instagram. Hasil penerapan metodologi kerangka kerja CRISP-DM dalam klasifikasi sentimen menggunakan algoritma *Random Forest* dengan koreksi ejaan menggunakan *Damerau-Levenshtein Distance* dengan penambahan *N-Gram* yang dikombinasikan dengan TF-IDF dapat meningkatkan akurasi dan memperbaiki kesalahan penulisan serta mengidentifikasi sentimen masyarakat mengenai IKN apakah menghasilkan kebermanfaatan bagi rakyat.

1.2 Perumusan Masalah

Pada bagian perumusan masalah, diuraikan mengenai permasalahan, metode penelitian serta pertanyaan penelitian yang diuraikan sebagai berikut:

1.2.1 Permasalahan

Pemindahan Ibu Kota Negara (IKN) dari Jakarta ke Kalimantan Timur menjadi isu yang banyak diperbincangkan di masyarakat, baik dari segi manfaat maupun tantangan. Topik ini telah memicu berbagai opini di media sosial, termasuk dukungan dan kritik terhadap kebijakan tersebut. Namun, dalam proses klasifikasi sentimen dari media sosial, sering kali ditemukan kesalahan penulisan oleh pengguna yang dapat mengurangi keakuratan analisis sentimen. Oleh karena itu, diperlukan metode yang mampu memperbaiki kesalahan penulisan serta meningkatkan akurasi klasifikasi sentimen terkait relokasi IKN.

1.2.2 Metode Usulan

Metode yang digunakan dalam penelitian ini melibatkan penggunaan *Damerau-Levenshtein Distance* sebagai metode koreksi ejaan untuk meningkatkan efisiensi dan akurasi. Algoritma klasifikasi yang digunakan untuk analisis sentimen adalah *Random Forest Classifier* menggunakan *N-Gram* yang dikombinasikan dengan TF-IDF untuk ekstraksi fitur dan *Information Gain* untuk seleksi fitur. Metode koreksi ejaan bertujuan memperbaiki kesalahan penulisan agar dapat meningkatkan keakuratan dalam analisis sentimen publik di media sosial. Metode kerangka kerja yang diusulkan dalam penelitian ini menggunakan CRISP-DM agar alur dari penelitian ini terencana dengan baik sesuai dengan tahapan yang ada dalam CRISP-DM.

1.2.3 Pertanyaan Penelitian

Berdasarkan uraian latar belakang, dapat diambil pertanyaan penelitian diantaranya sebagai berikut:

1. Apakah algoritma *Random Forest Classifier* menggunakan koreksi ejaan *Damerau-Levenshtein Distance* yang dikombinasikan *N-Gram* dan TF-IDF dengan *Information Gain* dapat menghasilkan analisis sentimen yang lebih akurat?
2. Apakah penghapusan koreksi ejaan dengan *Damerau-Levenshtein Distance* dan tokenisasi *N-Gram* mempengaruhi performa akurasi klasifikasi sentimen menggunakan *Random Forest*?
3. Bagaimana pengaruh dan keefektifan hasil perbandingan 3 teknik *Balancing Data* antara lain *Random Oversampling*, *Random Undersampling*, dan SMOTE ketika menggunakan dan tidak menggunakan koreksi ejaan *Damerau-Levenshtein Distance & N-Gram* dalam analisis sentimen?

1.3 Tujuan dan Manfaat

1.3.1 Tujuan

Berdasarkan rumusan masalah yang ada, dapat ditentukan tujuan dari penelitian ini diantaranya adalah sebagai berikut:

1. Mengukur keakuratan model analisis sentimen menggunakan algoritma *Random Forest Classifier* yang dikombinasikan dengan *Damerau-Levenshtein*

Distance, *N-Gram* dan TF-IDF untuk ekstraksi fitur, serta *Information Gain* untuk seleksi fitur.

2. Mengukur seberapa akurat model analisis sentimen menggunakan algoritma *Random Forest Classifier* jika tanpa menggunakan *N-Gram* dan *Damerau-Levenshtein Distance*.
3. Mengukur tingkat keefektifan dan pengaruh dari teknik *balancing data* yang digunakan pada masing-masing skenario.

1.3.2 Manfaat

Penelitian ini memiliki manfaat diantaranya sebagai berikut:

1. Penerapan *Damerau-Levenshtein Distance* dalam koreksi ejaan meningkatkan akurasi analisis sentimen publik terkait relokasi Ibu Kota Negara (IKN) dari media sosial.
2. Penelitian ini meningkatkan performa *Random Forest Classifier* melalui ekstraksi fitur menggunakan *N-Gram* dan TF-IDF serta seleksi fitur *Information Gain*, sehingga menghasilkan klasifikasi sentimen yang lebih akurat untuk data teks yang tidak terstruktur dari media sosial.
3. Penerapan metodologi kerangka kerja CRISP-DM memberikan pendekatan yang lebih terstruktur, terencana dan efektif.
4. Bagi pembaca dan penulis, menambah wawasan terkait pada bidang *text mining* dan *spelling correction* (perbaikan kata).

1.4 Batasan Masalah

Berikut batasan-batasan yang diberikan agar permasalahan tidak terlalu luas dan tidak menyimpang sehingga sesuai dengan tujuan penulis berikut ini:

1. Data yang digunakan untuk analisis sentimen terbatas pada platform media sosial TikTok (<https://t.tiktok.com/i18n/share/video/7403057706623307013/>), YouTube (<https://www.youtube.com/watch?v=gOqiOVzJsgE>) dan Instagram (https://www.instagram.com/p/C9_ar1gyuZd/), dalam rentang waktu Juni 2024 sampai dengan Desember 2024.
2. Fokus penelitian hanya pada penerapan *Damerau-Levenshtein Distance* sebagai metode koreksi ejaan.
3. Hanya menggunakan algoritma *Random Forest Classifier* dengan ekstraksi fitur *N-Gram* dikombinasikan dengan TF-IDF dan seleksi fitur *Information*

Gain yang digunakan untuk klasifikasi sentimen, tanpa perbandingan dengan algoritma lainnya.

1.5 Sistematika Penulisan Skripsi

Sistematika penulisan skripsi ini disusun dengan mengikuti aturan penulisan skripsi pada program studi S1 Sistem Informasi Universitas Trunojoyo Madura. Berikut merupakan metodologi penulisan proposal pada penelitian ini.

BAB I PENDAHULUAN

Bab ini menjelaskan terkait latar belakang, perumusan masalah, permasalahan penelitian, metode usulan, pertanyaan penelitian, tujuan dan manfaat penelitian serta batasan masalah yang dilakukan untuk klasifikasi sentimen relokasi Ibu Kota Nusantara, serta sistematika penulisan skripsi.

BAB II KAJIAN PUSTAKA

Teori penunjang ini berisi landasan teori yang terkait dengan relokasi Ibu Kota Nusantara, IKN, *Damerau-Levenshtein Distance*, *N-Gram*, TF-IDF, *Information Gain*, CRISP-DM, *Random Forest Classifier*. Bab ini juga mencantumkan beberapa penelitian terdahulu yang relevan dengan penelitian ini.

BAB III METODOLOGI PENELITIAN

Bab ini membahas terkait perancangan sistem menggunakan metodologi kerangka kerja CRISP-DM yang mencakup *bussiness understanding*, *data understanding*, *data preparation*, *modelling*, *evaluation* dan *deployment* dari proses klasifikasi sentimen relokasi Ibu Kota Nusantara menggunakan algoritma *Random Forest Classifier* serta perbaikan struktur kata dengan *Damerau-Levenshtein Distance* dengan kombinasi ekstraksi fitur *N-Gram* dan TF-IDF.

BAB IV HASIL DAN PEMBAHASAN

Bab ini membahas hasil pengolahan data, pengolahan dari hasil proses klasifikasi menggunakan algoritma *Random Forest Classifier* serta perbaikan struktur kata dengan *Damerau-Levenshtein Distance* dengan kombinasi ekstraksi fitur *N-Gram* dan TF-IDF.

BAB V PENUTUP

Bab ini berisi rangkuman kesimpulan dari hasil penelitian yang telah dilakukan. Selain itu, bab ini juga mencakup saran yang disampaikan untuk keperluan pengembangan penelitian di masa mendatang.

REFERENSI

Pada referensi berisi kumpulan rujukan jurnal, prosiding dari penelitian terdahulu yang digunakan untuk menyusun skripsi ini.

LAMPIRAN

Pada lampiran terdapat lampiran sampel data penelitian, kode lengkap program yang digunakan dalam penelitian ini.