

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi terjadi kian pesat hari demi hari. Seluruh sektor dan aspek kehidupan mengalami perkembangan dan modernisasi, tidak terkecuali sektor informasi dan komunikasi. Tidak hanya bagi perusahaan dan instansi pemerintah, kemajuan teknologi di sektor ini juga dapat dirasakan manfaatnya oleh seluruh lapisan, termasuk warga sipil. Baik anak sekolah, pemuda, hingga orang tua. Oleh karenanya, media konvensional telah banyak ditinggalkan, masyarakat beralih menggunakan *handphone* yang telah mendukung penggunaan internet untuk berkomunikasi. Pada tahun 2023, 78,19% dari total penduduk Indonesia atau sekitar 215,63 juta telah terhubung dengan internet, menurut data dari dataindonesia.id[1]. Penggunaan media sosial melalui internet memberikan kemudahan akses dalam bertukar informasi di manapun, dan kapanpun secara *real time* tanpa harus bertatap muka. Media sosial memberikan pengalaman berkomunikasi yang lebih menarik, karena memiliki fitur-fitur yang tidak ada di media konvensional seperti emoji, stiker, dan lain-lain. Hal ini menyebabkan terjadinya ikatan emosi yang lebih baik dalam berkomunikasi.

Lonjakan inovasi digital juga menyebabkan perkembangan pengguna yang terus meningkat[2]. Tidak hanya itu, *platform* media sosial pun semakin berkembang dan semakin beragam jenisnya. *Microblogging*, jejaring sosial, forum diskusi, *content sharing*, *instant messaging*, dan masih banyak lagi. Masing-masing memiliki keunggulan dan kekurangan tersendiri. Adapun beberapa aplikasi media sosial yang selalu ramai dibicarakan dan memiliki banyak pengguna adalah Instagram, TikTok, dan juga Youtube[3]. Instagram dan TikTok memiliki beberapa kesamaan. Keduanya merupakan *platform* bertukar pesan, sekaligus unggul dengan penyediaan konten foto serta video pendek. Adapun Youtube tetap menjadi *platform* teratas sebagai penyedia video panjang. Kemudahan akses informasi melalui media sosial juga memberikan kemudahan dalam mengemukakan pendapat tentang apapun secara bebas. Kemudahan ini juga dapat dimanfaatkan untuk mengetahui persepsi dari pengguna media sosial terhadap suatu hal, baik yang

sedang menjadi perbincangan hangat atau yang sudah berlalu. Mengenai berita, fenomena, individu, atau bahkan suatu kelompok. Salah satunya mengenai orang Madura. Orang Madura memiliki kebiasaan keluar dari kampung halamannya atau biasa dikenal dengan istilah merantau[4], sehingga menyebabkan orang Madura banyak dikenal oleh masyarakat luar daerah. Hal ini tentu membuka kemungkinan adanya opini masyarakat luar terhadap orang Madura di media sosial, baik positif atau negatif.

Teknologi terus berkembang dan akan terus dikembangkan setiap harinya. Penggunaan teknologi yang tepat akan memudahkan dalam mengesktrak teks dan mengolah informasi tersebut untuk berbagai keperluan bahkan bila dalam ukuran besar sekalipun. Ilmu ini dikenal dengan nama *text mining*. Proses *text mining* ini melibatkan berbagai metode, salah satunya *Natural Language Processing* (NLP). *Natural Language Processing* adalah bidang penelitian yang berkembang pesat dalam ilmu komputer dan kognitif[5] dan merupakan cabang dari kecerdasan buatan atau *artificial intelligence*. Penerapan *Natural Language Processing* dalam analisis sentimen akan memberikan hasil klasifikasi opini mengenai topik yang diinginkan. Analisis sentimen akan memberikan wawasan tentang tren dari persepsi publik dalam bentuk persentase opini positif dan negatif. Seiring dengan meningkatnya kebutuhan untuk menganalisis opini publik, para peneliti telah berupaya mengembangkan teknik dan algoritma otomatis yang akan memudahkan dalam proses analisis sentimen. Hasil ini nantinya dapat menjadi dasar pengambilan keputusan dalam menentukan langkah yang lebih baik.

Kemudahan dalam bertukar informasi di era teknologi saat ini menyebabkan interaksi terjadi dengan serba cepat. Berbalas pesan atau berkomentar di sebuah *paltform* dapat dilakukan hanya dalam sepersekian detik saja. Hal ini memungkinkan adanya kesalahan ketik atau *typo*. Namun, kesalahan tersebut merupakan hal yang wajar. Tahap pembersihan data atau *preprocessing* dalam *text mining* akan mengatasi kesalahan ini. Namun, meski begitu, penggunaan metode-metode perbaikan kata tentu akan memberikan hasil yang lebih optimal. Salah satu metode perbaikan kata yang seing kali digunakan dalam analisis sentimen adalah *Damerau-Levenshtein Distance*. Metode ini melakukan perbaikan kata dengan cara menghitung jarak perubahan minimal yang diperlukan untuk mengubah suatu kata

menjadi kata lainnya berdasarkan kamus atau panduan berisi kata yang benar. *Damerau-Levenshtein Distance* sendiri merupakan perkembangan dari *Levenshtein Distance*. Metode ini memiliki satu proses tambahan yang membedakannya dengan *Levenshtein Distance* yakni *transposition* (pertukaran posisi) dari tiga proses yang sudah ada yaitu *insertion* (penambahan), *deletion* (penghapusan) dan *substitution* (penggantian)[6]. Penelitian [6] juga menunjukkan adanya penambahan proses *transposition* ini membuat penggunaan *Damerau-Levenshtein Distance* menjadi lebih baik 2% dari *Levenshtein Distance* yang hanya menghasilkan akurasi 73% dan lebih unggul dari metode *Peter Norvig* yang hanya memiliki akurasi 69.09%. Perbandingan ini menunjukkan bahwa *Damerau-Levenshtein Distance* merupakan metode paling baik untuk melakukan mengoreksi kesalahan ejaan, sehingga *Damerau-Levenshtein Distance* dipilih untuk perbaikan kata pada penelitian ini.

Data yang telah dibersihkan akan dilanjutkan untuk diolah. *Naïve Bayes Classifier*, merupakan salah satu metode yang umum digunakan dalam analisis sentimen. *Naïve Bayes Classifier* merupakan sebuah algoritma yang berasas kepada teorema bayesian. Yakni, melakukan perhitungan kemungkinan kemunculan dari suatu kata yang ada pada setiap kelasnya, kemudian dilanjutkan dengan penentuan klasifikasinya. Adanya kata pembeda dari setiap kelas membuat penggunaan *Naïve Bayes* umumnya menghasilkan nilai akurasi yang tinggi. Selain itu, algoritma ini sering dipakai karena *Naïve Bayes* mampu mengklasifikasikan dokumen teks atau angka dengan sederhana dan memerlukan waktu komputasi yang terbilang cepat[7]. *Naïve Bayes Classifier* digunakan pada penelitian ini karena penelitian [8] menunjukkan akurasi *Naïve Bayes Classifier* bernilai 99,22%, paling tinggi dibandingkan algoritma KNN dan *Decision Tree* dengan akurasi 96,90% dan 37,21%. Namun, *Naïve Bayes Classifier* juga memiliki kelemahan. Asumsi independensi antar fitur bayesian dapat menurunkan performa dari algoritma ini, meski dalam kenyataan fitur-fitur tersebut seringkali memiliki keterkaitan. Penggunaan teknik ekstraksi fitur yang sesuai sangatlah penting dalam meningkatkan akurasi dan efisiensi waktu komputasi.

Penelitian [1] menunjukkan penggunaan ekstraksi fitur *N-Gram* dan pembobotan TF-IDF efektif dalam memberikan hasil yang lebih baik pada analisis sentimen. *N-Gram* substring sejumlah n kata atau karakter. *N-Gram* sendiri

memiliki fitur yang tidak terbatas, tetapi yang umum digunakan adalah *Unigram*, *Bigram*, dan *Trigram*. *Unigram*, *Bigram*, dan *Trigram* merupakan fitur-fitur dari *N-Gram* yakni berdasarkan satu kata, dua kata, dan tiga kata. *N-gram* mampu meningkatkan nilai presisi dan *recall*[1]. Selain itu, di antara ketiga fitur *N-gram*, penelitian [9] menunjukkan keunggulan *Bigram* daripada *Unigram* atau *Trigram* dengan akurasi sebesar 94,06%, presisi 90,71%, dan *recall* 91,73%. Penggunaan TF-IDF akan membantu dalam menemukan kata unik atau penting di dalam suatu dokumen sehingga akan meningkatkan performa dari analisis sentimen[10]. Oleh karena itu, kombinasi dari penggunaan *Naïve Bayes Classifier* dengan perbaikan kata *Damerau-Levenshtein Distance* serta ekstraksi fitur *N-Gram* ditambah pembobotan TF-IDF pada penelitian ini digunakan untuk menghasilkan tingkat akurasi klasifikasi analisis sentimen terhadap masyarakat Madura yang lebih optimal.

1.2 Perumusan Masalah

1.2.1 Permasalahan

Adanya persepsi masyarakat luar terhadap orang Madura yang tersebar di berbagai platform media sosial, baik dalam bentuk opini positif maupun negatif, menjadi tantangan dalam mengidentifikasi dan mengukur opini tersebut. Persepsi ini sulit diukur secara manual, terutama dengan banyaknya data yang tersedia di media sosial. Selain itu, kesalahan ejaan atau *typo* yang sering muncul pada komentar atau teks di media sosial dapat mempengaruhi akurasi hasil analisis sentimen. Kesalahan ejaan ini mempersulit algoritma untuk memahami makna dari teks yang dianalisis, sehingga diperlukan metode yang dapat memperbaiki ejaan dan mampu mengklasifikasikan opini publik dengan akurasi tinggi untuk memberikan gambaran yang jelas mengenai persepsi terhadap orang Madura.

1.2.2 Metode Usulan

Metode yang diusulkan melibatkan kombinasi beberapa teknik untuk meningkatkan akurasi analisis sentimen. Algoritma *Naïve Bayes Classifier* digunakan sebagai metode utama untuk klasifikasi teks, dikombinasikan dengan perbaikan kata *Damerau-Levenshtein Distance* dan fitur *Bigram Trigram* dari *N-Gram* serta pembobotan TF-IDF.

1.2.3 Pertanyaan Penelitian

Penelitian ini bertujuan untuk mendalami dan melakukan analisis sentimen terhadap masyarakat Madura. Secara jelas, penelitian ini memiliki tujuan dan urgensi sebagai berikut:

1. Bagaimana *Damerau-Levenshtein Distance* menangani kesalahan ejaan pada data sentimen menggunakan *Naïve Bayes Classifier*?
2. Bagaimana tingkat akurasi penggunaan *N-Gram* pada kombinasi *Naïve Bayes Classifier* dan *Damerau-Levenshtein Distance* dalam analisis sentimen terhadap masyarakat Madura?

1.3 Tujuan dan Manfaat

1.3.1 Tujuan Penelitian

Tujuan dalam penelitian ini yakni sebagai berikut:

1. Mengetahui penggunaan *Damerau-Levenshtein Distance* sebagai metode perbaikan kata dalam menangani kesalahan ejaan pada data sentimen.
2. Mengukur tingkat akurasi analisis sentimen dari penggunaan *N-Gram* pada kombinasi *Naïve Bayes Classifier* dan *Damerau-Levenshtein Distance*.

1.3.2 Manfaat Penelitian

Pada penelitian ini, manfaat yang diperoleh sebagai berikut:

1. Memberikan wawasan terkait penggunaan *Damerau-Levenshtein Distance* pada *Naïve Bayes Classifier* dalam menangani data sosial media yang rawan kesalahan ejaan dan *typo*.
2. Memberikan gambaran mengenai perbandingan performa *bigram* dan *trigram* dari *N-Gram* pada tingkat akurasi analisis sentimen terhadap masyarakat Madura.

1.4 Batasan Masalah

Adapun batasan masalah pada penelitian ini meliputi:

1. 766 data yang digunakan dalam penelitian ini diambil dari *platform* media sosial tertentu yakni Instagram, TikTok, serta YouTube.

2. Analisis hanya berfokus pada komentar teks berisi persepsi terkait masyarakat Madura, bentuk media (angka, emoji, dan lainnya) dan pandangan lain tidak dipertimbangkan.
3. Fokus penelitian ini adalah metode perbaikan ejaan atau struktur kata *Damerau-Levenshtein Distance* dan ekstraksi fitur *N-Gram*.
4. Algoritma yang digunakan adalah *Naïve Bayes Classifier* tanpa adanya perbandingan dengan algoritma lain.

1.5 Sistematika Penulisan

Sistematika penulisan yang digunakan terdiri dari lima bab sebagai berikut.

BAB I PENDAHULUAN

Bab ini berisi latar belakang, permasalahan, metode usulan, pertanyaan penelitian, tujuan penelitian, manfaat penelitian, batasan masalah, dan sistematika penulisan.

BAB II KAJIAN PUSTAKA

Pada bab ini akan dijelaskan teori-teori dari berbagai literatur yang akan digunakan sebagai dasar dalam penelitian ini.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan arsitektur sistem analisis sentimen, data penelitian, *labelling*, *preprocessing data*, perbaikan kata menggunakan *Damerau-Levenshtein Distance* dan ekstraksi fitur *N-Gram*, penjelasan pembobotan kata TF-IDF, klasifikasi *Naïve Bayes*, evaluasi model, dan skenario uji coba.

BAB IV HASIL DAN PEMBAHASAN

Pada bab ini menjelaskan mengenai hasil pengujian kombinasi algoritma *Naïve Bayes Classifier* dengan koreksi ejaan *Damerau-Levenshtein Distance* dan ekstraksi fitur *N-Gram* pada data komentar mengenai orang Madura.

BAB V PENUTUP

Bab ini berisi kesimpulan dari hasil pengujian serta saran untuk penelitian selanjutnya berdasarkan yang telah terjadi selama pengujian berlangsung.

REFERENSI

Bagian referensi berisi penelitian terkait yang menjadi sumber rujukan dalam penelitian.