

BAB I

PENDAHULUAN

1.1 Latar Belakang

Pada Pilpres Indonesia 2024, peran media sosial, khususnya Twitter, sangat signifikan sebagai platform diskusi politik. Sentimen masyarakat yang tercermin dari unggahan di Twitter dapat memberikan wawasan penting tentang opini publik terhadap pasangan calon presiden. Teknik undersampling digunakan agar model dapat mempelajari data secara seimbang dan menghasilkan prediksi yang akurat. Algoritma Naive Bayes dipilih karena kesederhanaan dan kehandalannya dalam tugas klasifikasi teks [1].

Perkembangan teknologi informasi dan komunikasi telah memberikan dampak signifikan terhadap cara manusia berinteraksi, salah satunya melalui media sosial. Twitter, sebagai salah satu platform media sosial yang populer di Indonesia, menjadi wadah bagi pengguna untuk menyampaikan opini, keluhan, atau apresiasi terhadap berbagai isu dan topik. Dalam konteks ini, analisis sentimen pada tweet berbahasa Indonesia memiliki peran penting untuk memahami persepsi masyarakat terhadap suatu topik tertentu [2].

Namun, tantangan utama dalam analisis sentimen adalah ketidakseimbangan distribusi kelas pada dataset. Data sentimen di Twitter sering kali didominasi oleh sentimen netral, sementara sentimen positif dan negatif memiliki proporsi yang jauh lebih kecil. Ketidakseimbangan ini dapat menyebabkan model klasifikasi memberikan bias terhadap kelas yang dominan, sehingga menurunkan akurasi dalam memprediksi sentimen lainnya [3]. Oleh karena itu, diperlukan metode untuk mengatasi permasalahan ketidakseimbangan data, seperti teknik undersampling yang berfokus pada penyamaan distribusi kelas [4].

Naive Bayes, sebagai salah satu algoritma pembelajaran mesin berbasis probabilistik, memiliki keunggulan dalam memproses data berukuran besar dan memberikan hasil yang cepat serta efisien. Kombinasi teknik undersampling dengan algoritma Naive Bayes memungkinkan pengembangan model yang lebih

akurat dalam mengklasifikasikan sentimen, meskipun data memiliki ketidakseimbangan [5]. Selain itu, penggunaan metode unsupervised clustering dalam pelabelan data juga dapat mempercepat proses klasifikasi awal, terutama pada dataset yang tidak memiliki label [6].

Penelitian ini bertujuan untuk menerapkan analisis sentimen pada tweet berbahasa Indonesia menggunakan algoritma Naive Bayes dengan teknik undersampling untuk mengatasi ketidakseimbangan kelas. Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan model klasifikasi sentimen yang lebih akurat, serta menjadi acuan bagi penelitian selanjutnya dalam bidang Natural Language Processing dan analisis sentiment[7].

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana teknik undersampling dapat membantu mengatasi ketidakseimbangan kelas dalam analisis sentimen pada data twitter bahasa indonesia terhadap pilpres indonesia 2024?
2. Bagaimana performa algoritma Naive Bayes dalam menganalisis sentimen pada tweet berbahasa Indonesia yang terdiri dari kategori positif, negatif, dan netral?

1.3 Batasan Masalah

Agar penelitian ini lebih terfokus dan dapat diselesaikan dengan tepat waktu, beberapa batasan masalah berikut diterapkan:

1. Data Twitter
Penelitian ini menggunakan data tweet berbahasa Indonesia yang diambil dari platform Twitter terkait topik Pilpres Indonesia 2024. Data yang digunakan adalah data tidak berlabel (unlabeled) yang diklasifikasikan ke dalam kategori sentimen dengan proses tertentu.
2. Klasifikasi Sentimen
Penelitian ini hanya memfokuskan klasifikasi sentimen menjadi tiga kategori, yaitu: positif, negatif, dan netral. Kategori ini dipilih untuk

memberikan gambaran umum tentang opini masyarakat terhadap Pilpres Indonesia 2024.

3. Algoritma Naive Bayes

Penelitian hanya menggunakan algoritma Naive Bayes untuk analisis sentimen karena kemampuannya yang efisien dalam menangani teks dan probabilitas meskipun pada dataset kecil dan terbatas.

4. Teknik Penyeimbangan Data

Teknik undersampling digunakan untuk menangani ketidakseimbangan kelas pada dataset sentimen. Teknik oversampling atau metode balancing lain tidak diterapkan dalam penelitian ini.

5. Preprocessing Data

Preprocessing data melibatkan langkah-langkah seperti tokenisasi, penghapusan stopword, stemming, dan representasi fitur menggunakan Term Frequency-Inverse Document Frequency (TF-IDF). Proses ini hanya dilakukan pada data mentah tanpa menambahkan elemen tambahan di luar teks tweet.

6. Evaluasi Model

Evaluasi model dilakukan dengan menggunakan metrik seperti akurasi, precision, recall, F1-score, dan confusion matrix. Analisis evaluasi hanya terbatas pada metrik ini tanpa memasukkan metrik tambahan.

7. Visualisasi Data

Visualisasi data dalam penelitian mencakup distribusi panjang teks setelah preprocessing, distribusi label sentimen setelah balancing data, word cloud untuk setiap kategori sentimen, dan matriks kesalahan (confusion matrix).

1.4 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah:

1. Mengembangkan model analisis sentimen menggunakan algoritma Naive Bayes untuk data Twitter terkait Pilpres Indonesia 2024.
2. Menganalisis pengaruh teknik undersampling dalam mengatasi ketidakseimbangan kelas pada dataset sentimen Twitter.

3. Mengevaluasi performa algoritma Naive Bayes dalam mengklasifikasikan sentimen menjadi kategori positif, negatif, dan netral pada data Twitter berbahasa Indonesia.
4. Memberikan kontribusi pada pengembangan model analisis sentimen berbasis Natural Language Processing (NLP) yang efisien dalam menangani data yang tidak seimbang.

Setiap tujuan ini diharapkan terukur dan tercapai melalui metode, analisis data, dan evaluasi hasil penelitian, yang akan disimpulkan di akhir penelitian.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Memberikan wawasan bagi pengelola Twitter dan pemangku kepentingan. penelitian ini dapat membantu pengelola platform media sosial, khususnya Twitter, dalam memahami sentimen publik terhadap topik tertentu yang sedang tren. Informasi ini dapat digunakan untuk memperbaiki kebijakan, strategi komunikasi, dan interaksi dengan pengguna, sehingga dapat meningkatkan pengalaman pengguna dan respons platform terhadap kebutuhan mereka.
2. Memberikan kontribusi bagi pengembangan algoritma klasifikasi sentiment. dengan penerapan teknik undersampling dan algoritma Naive Bayes, penelitian ini memberikan kontribusi dalam pengembangan model analisis sentimen yang lebih efisien, terutama dalam menangani dataset yang tidak seimbang. Teknik ini diharapkan dapat membantu peneliti dan pengembang aplikasi dalam mengatasi masalah ketidakseimbangan kelas yang sering ditemui dalam analisis data besar.
3. Menjadi referensi untuk penelitian selanjutnya dalam pengembangan analisis sentiment. penelitian ini diharapkan dapat memberikan dasar yang kuat bagi penelitian-penelitian selanjutnya dalam bidang analisis sentimen dan Natural Language Processing (NLP). Dengan mengembangkan model klasifikasi yang lebih akurat dan efisien, hasil dari penelitian ini dapat menjadi acuan untuk riset lebih lanjut yang berkaitan dengan optimisasi

algoritma pembelajaran mesin, serta penggunaan teknik lain untuk memperbaiki akurasi dalam analisis sentimen berbasis teks.

4. Memperkenalkan teknik baru untuk mengatasi masalah ketidakseimbangan data. teknik undersampling yang diterapkan dalam penelitian ini dapat menjadi solusi untuk masalah ketidakseimbangan kelas yang seringkali mengganggu kinerja model klasifikasi. Hasil dari penelitian ini dapat membantu praktisi dan akademisi dalam memahami cara terbaik untuk menangani ketidakseimbangan data, terutama pada analisis sentimen yang memerlukan prediksi terhadap berbagai kategori sentimen (positif, negatif, netral).

1.6 Sistematika Penulisan

Penulisan skripsi ini disusun dengan sistematika sebagai berikut:

BAB I PENDAHULUAN

Bab ini berisi latar belakang penelitian yang menjelaskan alasan mengapa topik analisis sentimen pada tweet berbahasa Indonesia dipilih. Di dalamnya juga terdapat rumusan masalah yang menggambarkan permasalahan yang akan dipecahkan dalam penelitian ini, tujuan dan manfaat penelitian, batasan masalah yang akan diteliti, serta sistematika penulisan skripsi. Bab ini memberikan gambaran umum mengenai fokus penelitian dan memberikan konteks tentang pentingnya penelitian ini dilakukan.

BAB II TINJAUAN PUSTAKA

Bab ini menyajikan landasan teori yang digunakan dalam penelitian, termasuk konsep dasar analisis sentimen, algoritma Naive Bayes, serta teknik undersampling untuk mengatasi ketidakseimbangan kelas dalam dataset. Di bab ini juga akan dibahas mengenai teknik unsupervised clustering yang digunakan dalam pelabelan data, serta tinjauan pustaka dari penelitian-penelitian terdahulu yang relevan dengan topik penelitian. Tujuan bab ini adalah untuk memberikan fondasi teoretis yang mendasari penerapan metode dalam penelitian ini.

BAB III METODE PENELITIAN

Bab ini menguraikan metode yang digunakan dalam penelitian, mulai dari objek

dan jenis data, sumber data tweet bahasa Indonesia yang digunakan, serta teknik pengumpulan data. Selain itu, dijelaskan secara rinci tahapan dalam pemrosesan data yang meliputi pembersihan data, tokenisasi, stemming, dan pelabelan data menggunakan teknik unsupervised clustering. Proses pembuatan model klasifikasi menggunakan Naive Bayes dan teknik undersampling juga akan dibahas dalam bab ini, termasuk metode evaluasi model menggunakan metrik akurasi, precision, recall, dan F1-score. Bab ini bertujuan untuk menjelaskan tahapan sistematis yang ditempuh untuk mencapai tujuan penelitian.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil dari analisis sentimen yang dilakukan pada dataset tweet berbahasa Indonesia, termasuk distribusi sentimen positif, negatif, dan netral. Hasil evaluasi model Naive Bayes setelah diterapkan teknik undersampling juga akan dianalisis di sini. Pembahasan mendalam dilakukan mengenai performa model, termasuk penggunaan metrik evaluasi seperti akurasi, precision, recall, dan F1-score. Pembahasan juga mencakup visualisasi data dan analisis confusion matrix untuk mengidentifikasi kesalahan model dan peluang perbaikan.

BAB V PENUTUP

Bab ini berisi kesimpulan dari penelitian berdasarkan hasil yang telah dianalisis dan dibahas pada bab sebelumnya. Kesimpulan akan mencakup pencapaian tujuan penelitian dan implikasi dari hasil yang diperoleh. Di bab ini juga disertakan saran-saran untuk penelitian selanjutnya yang dapat memperbaiki kekurangan atau mengembangkan metode yang telah diterapkan.