

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi dan komunikasi yang pesat telah mendorong masyarakat untuk lebih aktif menyampaikan pendapat melalui platform digital, salah satunya adalah YouTube [1]. YouTube tidak hanya menjadi media hiburan, tetapi juga sarana penyebaran informasi dan diskusi publik, termasuk dalam ranah politik [2]. Salah satu konten politik yang menarik perhatian publik adalah video berjudul "Presiden Prabowo Menjawab", di mana Presiden Prabowo Subianto menjawab berbagai pertanyaan dari jurnalis terkait program kerja, isu keamanan, ekonomi, hingga kebijakan luar negeri. Video tersebut memancing beragam respons dari masyarakat yang dapat diamati melalui kolom komentar. Respons tersebut mencerminkan opini publik yang sangat bermanfaat untuk dianalisis secara sistematis. Salah satu pendekatan yang dapat digunakan untuk memahami opini tersebut adalah analisis sentimen.

Analisis sentimen merupakan bagian dari pemrosesan bahasa alami (*Natural Language Processing* / NLP) yang berfokus pada proses mengenali, menilai, dan menginterpretasikan opini atau sentimen yang terkandung dalam suatu teks [3]. Seiring dengan meningkatnya penggunaan media sosial dan berkembangnya era digital, analisis sentimen menjadi semakin penting karena dapat memberikan pemahaman mengenai persepsi dan tanggapan pengguna terhadap suatu produk, layanan, maupun isu tertentu [4]. Dalam proses analisis sentimen, terdapat berbagai metode yang dapat digunakan untuk mengklasifikasikan data teks berdasarkan sentimen yang dikandungnya. Beberapa metode tersebut antara lain K-Nearest Neighbor (KNN), Decision Tree, Naïve Bayes, Random Forest, serta algoritma klasifikasi lainnya. Dalam penelitian ini, algoritma yang digunakan adalah Multinomial Naïve Bayes. Algoritma ini memiliki asumsi bahwa setiap fitur atau kata dalam data bersifat independen, meskipun pada kenyataannya kata-kata dalam kalimat saling berkaitan [5] [6] [7]. Namun, dengan kesederhanaan dan efisiensinya, algoritma

ini tetap terbukti mampu memberikan hasil yang baik dalam klasifikasi teks [8] [9] [10] [11] [12].

Pemilihan algoritma Naïve Bayes dalam penelitian ini didasarkan pada temuan dari beberapa studi terdahulu yang menunjukkan performa unggulnya dalam tugas klasifikasi teks. penelitian oleh Desyani, seleksi fitur terbukti signifikan dalam meningkatkan kinerja klasifikasi menggunakan algoritma Naive Bayes pada data medis. Studi tersebut menunjukkan bahwa metode SelectKBest menghasilkan performa tertinggi dalam memprediksi kebutuhan operasi caesar, dengan peningkatan akurasi dan AUC dibandingkan tanpa seleksi fitur [13]. Hasil yang sama dari penelitian Rochim, analisis sentimen menggunakan algoritma Naive Bayes tanpa seleksi fitur pada dataset lockdown COVID-19 menghasilkan akurasi rata-rata 63,2%, sedangkan dengan penerapan Chi-Square sebagai metode seleksi fitur, akurasi meningkat menjadi 65%. Studi komparatif ini menunjukkan bahwa seleksi fitur Chi-Square secara konsisten mengungguli Information Gain dengan akurasi 64,2% dan model baseline tanpa seleksi fitur, karena kemampuannya menyaring fitur yang lebih relevan dengan kategori sentimen [14]. Dan kemudian dengan penelitian Saffiri, akurasi klasifikasi Naive Bayes pada data customer churn yang tidak seimbang hanya mencapai 73%. Penerapan teknik SMOTE untuk menangani ketidakseimbangan kelas berhasil meningkatkan akurasi menjadi 74,9%, sementara kombinasi SMOTE dan Genetic Algorithm untuk seleksi fitur lebih lanjut berhasil mendongkrak akurasi hingga 80,9%. Hasil ini membuktikan bahwa penanganan class imbalance dan seleksi fitur merupakan pendekatan kritis untuk mengoptimalkan kinerja model Naive Bayes [15].

Dari identifikasi gap tersebut, penelitian ini mengusulkan sebuah kerangka kerja komprehensif untuk mengoptimalkan kinerja klasifikasi Multinomial Naïve Bayes (MNB) dengan menangani dua tantangan utama, yaitu dimensionalitas data yang tinggi dan ketidakseimbangan distribusi kelas. Pada tahap pra-pemrosesan, teknik seleksi fitur menggunakan Chi-Square diterapkan untuk mereduksi dimensi data dengan secara statistik memilih dan mempertahankan fitur-fitur yang paling informatif dan relevan terhadap variabel

target, sehingga tidak hanya meningkatkan efisiensi komputasi dengan menyederhanakan model tetapi juga berpotensi meningkatkan akurasi dengan mengurangi noise dari fitur yang tidak signifikan [16]. Secara paralel, untuk mengatasi masalah ketidakseimbangan kelas yang dapat menyebabkan model bias terhadap kelas mayoritas, teknik Synthetic Minority Over-sampling Technique (SMOTE) diintegrasikan untuk menciptakan sampel sintetis yang realistis bagi kelas minoritas dengan melakukan interpolasi pada ruang fitur, sehingga distribusi kelas menjadi lebih seimbang dan model MNB dapat belajar karakteristik setiap kelas dengan lebih baik [17] [18]. Sinergi dari kedua metode ini diharapkan mampu menghasilkan model MNB yang lebih robust, akurat, dan generalis dalam melakukan klasifikasi.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk melakukan analisis sentimen terhadap komentar pada video "Presiden Prabowo Menjawab" menggunakan algoritma Multinomial Naïve Bayes. Untuk meningkatkan performa model, dilakukan seleksi fitur menggunakan metode Chi-Square dan penyeimbangan data menggunakan SMOTE. Kombinasi ini diharapkan mampu mengurangi beban komputasi sekaligus meningkatkan akurasi dalam klasifikasi sentimen komentar publik terhadap tokoh politik di era digital.

1.2 Perumusan Masalah

Berdasarkan penjelasan latar belakang yang telah disampaikan di atas, rumusan masalah dapat dijabarkan sebagai berikut.

1.2.1 Permasalahan

Berdasarkan latar belakang yang telah dijelaskan, permasalahan yang dihadapi adalah banyaknya komentar pada YouTube yang cenderung bernuansa negatif, sehingga diperlukan metode yang efektif untuk menyeimbangkan sentimen tersebut. Algoritma Naïve Bayes dipilih karena merupakan algoritma yang relatif sederhana dengan akurasi yang tinggi. Namun demikian, algoritma ini masih memiliki kelemahan, terutama dalam hal waktu komputasi yang cukup lama. Hal ini disebabkan oleh keterbatasannya dalam menangani fitur yang

kurang relevan, sehingga diperlukan kombinasi dengan teknik seleksi fitur yang lebih tepat untuk meningkatkan efisiensi dan akurasi.

1.2.2 Metode Usulan

Solusi yang diusulkan dalam penelitian ini adalah penerapan teknik oversampling data dengan menggunakan metode SMOTE untuk menyeimbangkan dataset, yang kemudian diproses menggunakan algoritma Naïve Bayes. Algoritma ini juga dikombinasikan dengan seleksi fitur Chi-Square untuk mengurangi fitur-fitur yang kurang relevan, sehingga dapat mempercepat waktu komputasi.

1.2.3 Pertanyaan Penelitian

Dalam penelitian ini terdapat pertanyaan yang diperoleh dari permasalahan dan metode usulan yang telah disebutkan diatas ialah berapa besar pengaruh peningkatan penggunaan chi square serta smote dalam menganalisis sebuah komentar dengan menggunakan multinomial naïve bayes ?

1.3 Tujuan dan Manfaat

Tujuan dan Manfaat dalam penelitian ini sebagai berikut :

1.3.1 Tujuan Penelitian

Pada penelitian ini terdapat beberapa tujuan yang ingin dicapai, diantaranya sebagai berikut :

- a. Untuk mengetahui berapa besar pengaruh peningkatan penggunaan seleksi fitur chi-square serta penggunaan smote dalam menganalisis sebuah komentar dengan menggunakan multinomial naïve bayes
- b. Membuat Aplikasi klasifikasi sentiment analisis komentar youtube terhadap video “Presiden Prabowo Menjawab”

1.3.2 Manfaat Penelitian

Pada penelitian ini terdapat beberapa manfaat yang dapat diperoleh, diantaranya sebagai berikut:

- a. Dapat digunakan untuk menganalisis seberapa efektif metode SMOTE dan seleksi fitur Chi-Square terhadap peningkatan performa algoritma Naïve Bayes yang digunakan dalam pengklasifikasian sentimen ulasan.
- b. Dapat digunakan sebagai bahan rujukan untuk penelitian selanjutnya.

1.4 Batasan Masalah

Dalam penelitian ini ada batasan-batasan yang diterapkan, diantaranya sebagai berikut :

1. Dataset yang digunakan yakni data komentar youtube pada beberapa channel youtube seperti Najwa Shihab, Narasi dan Kompas TV.
2. Dataset yang digunakan memiliki 2 label yaitu label positif dan label negative.
3. Metode yang digunakan pada penelitian ini yaitu metode Multinomial Naïve Bayes dengan seleksi fitur chi square dan smote. Tidak dengan algoritma lain.
4. Teknik Crawling Pada penelitian ini menggunakan code program python dengan selenium

1.5 Sistematika Proposal

BAB I Pendahuluan :

Bab ini berisi latar belakang penelitian mengenai analisis sentimen video "Presiden Prabowo Menjawab" di YouTube menggunakan algoritma Multinomial Naïve Bayes (MNB). Dibahas permasalahan utama seperti ketidakseimbangan kelas dan dimensionalitas data tinggi yang mempengaruhi akurasi MNB. Diperkenalkan solusi kombinasi SMOTE dan Chi-Square Feature Selection untuk meningkatkan performa model. Bab ini juga mencakup rumusan masalah, tujuan, manfaat, dan batasan penelitian.

BAB II Tinjauan Pustaka :

Bab ini menyajikan landasan teori yang meliputi konsep analisis sentimen, teknik web scraping untuk pengumpulan data dari YouTube, dan algoritma

Multinomial Naïve Bayes. Juga dijelaskan teori pendukung mengenai metode SMOTE untuk handling class imbalance dan Chi-Square untuk seleksi fitur, serta penelitian terdahulu yang relevan.

BAB III Metode Penelitian :

Bab ini menjelaskan metodologi penelitian yang digunakan, dimulai dari pengumpulan data komentar YouTube, preprocessing data, labeling sentimen, penerapan SMOTE dan Chi-Square, pembangunan model MNB, hingga evaluasi model menggunakan metrik akurasi, presisi, recall, dan F1-score.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil implementasi metode yang diusulkan dan analisis terhadap temuan penelitian. Dibandingkan performa model MNB dasar dengan model yang telah dioptimasi menggunakan SMOTE dan Chi-Square. Hasil evaluasi dibahas secara mendalam untuk melihat efektivitas metode yang diusulkan.

BAB V KESIMPULAN

Bab ini berisi kesimpulan akhir dari seluruh penelitian yang telah dilakukan, menjawab pertanyaan penelitian, serta menyampaikan saran dan rekomendasi untuk pengembangan penelitian sejenis di masa mendatang.