

BAB I

PENDAHULUAN

1.1 Latar Belakang

Di era digital saat ini, analisis sentimen menjadi pendekatan penting dalam memahami opini publik terhadap berbagai isu, termasuk isu-isu sosial dan pemerintahan seperti korupsi. Korupsi merupakan permasalahan serius yang dapat merusak tatanan hukum, ekonomi, dan kepercayaan masyarakat terhadap institusi negara. Salah satu kasus yang menarik perhatian publik adalah dugaan isu korupsi yang terjadi pada distribusi bahan bakar Pertamina, sebuah produk yang secara luas digunakan oleh masyarakat Indonesia. Reaksi dan opini publik terhadap isu ini tersebar di berbagai platform digital, sehingga penting untuk dilakukan analisis sentimen guna mengetahui kecenderungan opini masyarakat terhadap kasus tersebut[1].

Permasalahan dugaan isu korupsi Pertamina memiliki dampak signifikan terhadap tingkat kepercayaan masyarakat terhadap Badan Usaha Milik Negara (BUMN). Meskipun hingga saat pengumpulan data belum terdapat putusan hukum yang bersifat final, dinamika opini publik menunjukkan peningkatan yang pesat. Kondisi ini menegaskan pentingnya dilakukan pemetaan persepsi masyarakat melalui analisis sentimen untuk memperoleh gambaran yang objektif terkait pandangan publik serta implikasinya terhadap citra institusi negara.

Analisis sentimen merupakan proses pemrosesan dan klasifikasi teks secara otomatis untuk mengidentifikasi sikap atau opini seseorang terhadap suatu topik [2]. Proses ini termasuk dalam bidang *text mining*, yang memanfaatkan metode seperti *Natural Language Processing* (NLP), *information retrieval*, dan *machine learning* untuk mengolah data teks dalam skala besar. Salah satu algoritma klasifikasi yang banyak digunakan dalam analisis sentimen adalah *Support Vector Machine* (SVM), yang memiliki kemampuan tinggi dalam menemukan pola dari data teks dan mengelompokkannya ke dalam kelas sentimen, seperti positif dan negatif [3]. SVM bekerja dengan mencari *hyperplane* terbaik yang memisahkan antar kelas dalam ruang berdimensi tinggi, dan dalam penelitian ini digunakan bersamaan dengan metode *Term Frequency-Inverse Document Frequency* (TF-IDF) untuk mengekstraksi fitur penting dari teks ulasan publik.

Keberhasilan model klasifikasi seperti SVM sangat bergantung pada kualitas dan distribusi data yang digunakan. Salah satu tantangan utama yang sering dihadapi adalah ketidakseimbangan data (*imbalanced dataset*), yaitu kondisi ketika jumlah data dalam satu kelas jauh lebih banyak dibandingkan kelas lainnya. Ketidakseimbangan ini dapat menyebabkan model lebih cenderung untuk memprediksi kelas mayoritas dan mengabaikan kelas minoritas. Dalam konteks penelitian ini, dataset yang digunakan menunjukkan ketidakseimbangan kelas, dengan total 900 data berlabel sentimen negatif dan hanya 120 data berlabel sentimen positif. Proporsi ini menunjukkan bahwa 88.24 % data merupakan opini negatif, sementara hanya 11,76% yang bersentimen positif. Ketidakseimbangan kelas ini berpotensi mengakibatkan bias dalam proses klasifikasi, di mana model lebih sering mengklasifikasikan teks sebagai negatif, sehingga hasil analisis menjadi kurang representatif dan tidak mencerminkan opini publik secara menyeluruh.

Dalam mengatasi permasalahan ketidakseimbangan data ini, penelitian ini secara spesifik memilih *text augmentation* dan SMOTE (*Synthetic Minority Over-sampling Technique*). *Text augmentation* dipilih karena kemampuannya untuk memperluas dataset pelatihan dengan membuat variasi teks baru tanpa mengubah makna asli, yang sangat relevan untuk data tekstual[4]. Dalam konteks ini, dua metode *text augmentation* yang diterapkan adalah *synonym replacement* dan *back translation*. *Synonym replacement* dilakukan dengan mengganti kata-kata dalam kalimat dengan sinonimnya, sedangkan *Back Translation* dilakukan dengan menerjemahkan kalimat ke bahasa asing, lalu mengembalikannya ke bahasa asal untuk menghasilkan struktur kalimat yang berbeda namun tetap mempertahankan makna inti[5].

Selain mengatasi permasalahan ketidakseimbangan data yang digunakan teknik augmentasi, SMOTE dipilih sebagai teknik *oversampling* sintesis yang efektif untuk data terstruktur, dengan menghasilkan sampel baru berdasarkan interpolasi data minoritas yang ada. Pemilihan kedua metode ini didasarkan pada efektivitasnya dalam konteks data tekstual dan numerik (setelah TF-IDF), serta untuk membandingkan pendekatan berbasis manipulasi teks langsung (*text augmentation*) dengan pendekatan berbasis fitur numerik SMOTE dalam seimban

kelas[6].

Pentingnya penelitian mengenai analisis sentimen telah mendorong peneliti untuk melakukan studi lebih lanjut. Penelitian yang dilakukan oleh[7] bertujuan untuk menilai efektivitas analisis opini masyarakat. Data yang digunakan terdiri dari 2000 data *tweet* yang dikumpulkan melalui *web scraping* dari sosial media X, dengan hasil akurasi sebesar 87,6%, presisi 96,53%, dan *recall* 85,67%.

Penelitian berikutnya[8] bertujuan untuk mengevaluasi kinerja metode *Support Vector Machine* (SVM) dengan kernel *linier* dan RBF dalam analisis sentimen. Data dikumpulkan melalui proses *crawling* di Twitter menggunakan kata kunci 'Boikot pro Israel', menghasilkan 2.600 *tweet*. Proses pelabelan dilakukan secara manual, dengan 1.560 *tweet* dikategorikan sebagai dukungan dan 1.040 sebagai nondukungan. Hasil implementasi model menunjukkan bahwa kernel *linier* mencapai akurasi 92,5%, sedangkan kernel RBF memperoleh akurasi 91,92%. Penelitian ini mengungkapkan bahwa sentimen positif lebih dominan, dengan 1.560 *tweet* mendukung dan 1.040 *tweet* menentang gerakan boikot.

Penelitian selanjutnya yang dilakukan oleh[9] telah melakukan penelitian mengenai penerapan teknik augmentasi data untuk mengatasi ketidakseimbangan data pada klasifikasi teks berbahasa Indonesia. Fokus penelitian ini adalah penggunaan teknik augmentasi data untuk meningkatkan keseimbangan data dalam klasifikasi teks. Metode yang digunakan meliputi *Oversampling*, *Synonym Replacement*, dan *Back Translation* untuk memperluas data pelatihan dan meningkatkan performa model klasifikasi. Hasil penelitian menunjukkan bahwa penerapan teknik *Back Translation* mampu meningkatkan skor F1 untuk label minoritas sebesar 8%, sedangkan teknik *Synonym Replacement* memberikan peningkatan sebesar 5%. Penelitian ini memberikan kontribusi penting dalam pengolahan teks berbahasa Indonesia, khususnya dalam meningkatkan akurasi dan keseimbangan data pada kasus deteksi *clickbait* dan komentar ujaran kebencian pada berita daring.

Penelitian yang dilakukan oleh[10] menjadi salah satu referensi penting dalam mengatasi permasalahan ketidakseimbangan data pada analisis sentimen. Studi tersebut memfokuskan diri pada klasifikasi sentimen masyarakat terhadap program vaksinasi Covid-19 dengan menggunakan algoritma *Support Vector*

Machine (SVM) dan *Naive Bayes*, serta menerapkan metode *oversampling Synthetic Minority Over-sampling Technique* (SMOTE) untuk menyeimbangkan distribusi kelas. Hasil penelitian menunjukkan bahwa penerapan SMOTE mampu meningkatkan performa model, dengan akurasi mencapai 70,51%. Temuan ini memperkuat urgensi penggunaan metode penanganan data tidak seimbang guna meningkatkan efektivitas klasifikasi dalam studi analisis sentimen.

Penelitian yang dilakukan oleh [11] memberikan kontribusi signifikan dalam menunjukkan potensi teknik augmentasi data dalam meningkatkan kinerja klasifikasi teks. Dalam studi tersebut, peneliti memperkenalkan metode *Easy Data Augmentation* (EDA) sebagai pendekatan untuk memperluas data pelatihan, terutama pada dataset berukuran kecil. Teknik augmentasi yang diterapkan meliputi *Synonym Replacement*, *Back Translation*, dan *Function Word Deletion*, serta diuji menggunakan berbagai algoritma seperti *Random Forest*, *Logistic Regression*, dan *Naive Bayes*. Hasil penelitian menunjukkan bahwa teknik *Back Translation* memberikan kontribusi akurasi tertinggi, yaitu sebesar 90,60% saat digunakan bersama algoritma *Random Forest*, sementara *Synonym Replacement* mencapai akurasi 82,77%. Temuan ini mengindikasikan bahwa *text augmentation* merupakan strategi yang efektif untuk meningkatkan akurasi klasifikasi, khususnya dalam kondisi keterbatasan data.

Dengan penelitian ini, dapat diperoleh pemahaman yang lebih dalam mengenai efektivitas penggunaan teknik augmentasi data dalam meningkatkan kinerja model analisis sentimen terhadap isu-isu sosial seperti kasus korupsi Pertamina. Penelitian ini juga dapat memberikan kontribusi terhadap pengembangan strategi *preprocessing* data teks dalam pengolahan bahasa alami, khususnya dalam konteks data yang terbatas dan tidak seimbang.

1.2 Perumusan Masalah

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

1.2.1 Permasalahan

Data sentimen mengenai isu korupsi Pertamina di media sosial X memiliki distribusi yang tidak seimbang antara sentimen positif dan negatif. Keterbatasan jumlah data latih dapat mempengaruhi kinerja model dalam mengklasifikasikan

sentimen secara akurat. Penggunaan metode klasifikasi SVM tanpa teknik augmentasi dapat menghasilkan model yang kurang optimal dalam mengenali pola sentimen.

1.2.2 Metode Usulan

Pada penelitian ini memanfaatkan dua metode yaitu teknik *text augmentation* dan teknik SMOTE untuk memperbaiki kualitas data latih, diterapkan TF-IDF untuk ekstraksi fitur serta meningkatkan kemampuan generalisasi model *Support Vector Machine* (SVM).

1.2.3 Pertanyaan Penelitian

Berdasarkan latar belakang yang telah diuraikan maka dapat diketahui rumusan masalah sebagai berikut :

1. Berapa nilai akurasi yang diperoleh dengan menggunakan *Support Vector Machine* (SVM) dengan *oversampling* SMOTE dibandingkan menggunakan *text augmentation*?
2. Bagaimana pengaruh penggunaan kernel yang berbeda (Linear dan RBF) dan optimasi parameter (C dan *Gamma*) terhadap nilai akurasi yang diperoleh pada skenario penyeimbangan data?

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. Untuk mengetahui hasil nilai akurasi yang diperoleh isu korupsi Pmertamina menggunakan *Support Vector Machine* dengan *oversampling* SMOTE dibandingkan menggunakan *text augmentation*.
2. Untuk mengetahui nilai akurasi dari penggunaan kernel (Linear dan RBF) dan optimasi parameter (C dan *Gamma*) yang diperoleh pada skenario penyeimbangan data.

1.3.2 Manfaat

Penelitian ini diharapkan dapat memberikan gambaran penerapan *Support Vector Machine* (SVM) untuk analisis sentimen isu korupsi Pertamina serta menjadi referensi perbandingan efektivitas dua metode penyeimbangan data, yaitu

text augmentation (*Synonym Replacement* dan *Back Translation*) dan SMOTE, dalam meningkatkan performa klasifikasi data teks.

1.4 Batasan Masalah

Penelitian ini menetapkan batasan masalah untuk memastikan fokus dan kejelasan dalam pembahasan yang dilakukan. Adapun batasan masalah dari penelitian ini adalah sebagai berikut:

1. Sumber data yang digunakan dalam penelitian ini berasal dari hasil *crawling* data sosial media X dengan 5 kata kunci yaitu 'kasus korupsi pertamina', 'pertamax pertalite', 'pertamax oplosan', 'korupsi pertamina', dan 'pertamina sarang koruptor' mulai tanggal 25-28 Februari 2025. Pemilihan kelima kata kunci ini didasarkan pada relevansinya yang tinggi dengan isu korupsi Pertamina yang sedang hangat diperbincangkan pada periode tersebut, serta untuk mencakup berbagai aspek perbincangan mulai dari dugaan korupsi hingga produk terkait dan sentimen negatif yang eksplisit.
2. Ekstraksi fitur menggunakan metode TF-IDF (*Term Frequency - Inverse Document Frequency*).
3. Pelabelan data secara manual.
4. Algoritma yang digunakan untuk klasifikasi adalah *Support Vector Machine* (SVM).
5. Klasifikasi sentiment dibagi menjadi dua, yaitu sentimen positif dan negatif.
6. Jumlah data yang digunakan dalam penelitian ini sebanyak 1.020 ulasan.
7. Penyeimbangan data ulasan dan menambah variasi dalam kata menggunakan dua metode yaitu teknik pertama *Text Augmentation* dengan teknik *Synonym Replacement* dan *Back Translation*. Dan teknik kedua menggunakan SMOTE.
8. Pada penelitian ini untuk analisis dan pemrosesan data menggunakan *Python* dan *Google Collab*.

1.5 Sistematika Penulisan

Sistematika penulisan skripsi ini penulis susun dengan mengikuti aturan penulisan skripsi pada program studi Sistem Informasi Universitas Trunojoyo Madura. Berikut metodologi penulisan skripsi pada penelitian ini.

BAB I PENDAHULUAN

Bab ini menjelaskan mengenai latar belakang, perumusan masalah, permasalahan, metode usulan, pertanyaan penelitian, tujuan dan manfaat penelitian, batasan masalah, serta sistematika penulisan terkait penelitian yang dilakukan.

BAB II TINJAUAN PUSTAKA

Bab ini membahas mengenai landasan teori terkait korupsi bensin pertamax, serta metode menggunakan *Support Vector Machine*. Bab ini juga merinci proses penelitian, mulai dari teknik *preprocessing*, teknik *augmentation*, SMOTE, pelabelan, ekstraksi fitur hingga evaluasi model, serta mencakup penelitian terdahulu yang relevan.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan mengenai metode penelitian yang dilakukan, yaitu mengenai dataset yang digunakan, arsitektur sistem, serta skenario-skenario yang di uji coba.

BAB IV HASIL DAN PEMBAHASAN

Bab ini membahas hasil implementasi metode *Support Vector Machine* (SVM) untuk klasifikasi sentimen pada data isu Korupsi Pertamina. Pengujian dilakukan dengan dua skenario berbeda yang membandingkan teknik penyeimbangan data *text augmentation* dan SMOTE. Evaluasi dilakukan berdasarkan akurasi, presisi, recall, dan F1-score, serta analisis pengaruh kernel dan parameter terhadap performa model.

BAB V PENUTUP

Membahas kesimpulan yang didapatkan selama melakukan penelitian beserta saran – saran untuk penelitian selanjutnya

LAMPIRAN

Berisi memuat *source code* utama yang digunakan dalam penelitian, termasuk proses *preprocessing data*, penerapan metode *Support Vector Machine* (SVM), teknik penyeimbangan data (*text augmentation* dan SMOTE), ekstraksi fitur dengan TF-IDF, serta proses pelatihan dan evaluasi model.