BAB I

PENDAHULUAN

I.1 Latar Belakang

Permasalahan *Imbalanced* data menjadi hambatan signifikan dalam analisis sentimen karena menyebabkan model klasifikasi bias terhadap kelas mayoritas, sehingga gagal mengenali opini yang bernilai dari kelas minoritas (Rifka Noor Ikhsan dkk., 2024). Konteks e-commerce seperti Tokopedia, kegagalan ini berisiko mengabaikan keluhan pelanggan yang krusial. Ulasan pelanggan sangat penting karena ulasan yang mengandung sentimen negatif sering kali merepresentasikan pengalaman buruk yang membutuhkan perhatian segera dari pihak pengembang atau manajemen aplikasi. Jika keluhan tersebut tidak terdeteksi akibat ketidakseimbangan data, maka dapat menurunkan kualitas layanan dan berdampak pada kepercayaan serta loyalitas pengguna.

Tokopedia sebagai salah satu platform e-commerce terbesar di Indonesia, memiliki jumlah pengguna yang sangat besar dan volume ulasan yang terus bertambah. Ulasan-ulasan ini mencerminkan persepsi konsumen terhadap layanan Tokopedia dan menjadi sumber data penting untuk pengambilan keputusan strategis (Maulana dkk., 2023). Pengolahan data ulasan secara akurat dengan mengatasi ketidakseimbangan data menjadi sangat krusial dalam menjamin sistem analisis sentimen yang adil dan efektif.

Salah satu pendekatan yang umum digunakan untuk mengatasi masalah ini adalah teknik *oversampling*, yakni dengan meningkatkan jumlah data pada kelas minoritas agar distribusi kelas menjadi seimbang (Bej dkk., 2021). Berbagai metode

oversampling telah dikembangkan, termasuk *Random Oversampling*, SMOTE, dan *Borderline-SMOTE* (Belhaouari dkk., 2024). Beberapa studi sebelumnya telah dilakukan untuk mengatasi permasalahan ketidakseimbangan data (*Imbalanced* data). Misalnya, penelitian (E. Fitri dkk., 2020) menerapkan metode SMOTE dengan tiga algoritma klasifikasi, yaitu *Naïve Bayes*, *Random Forest*, dan *Support Vector Machine*. Penelitian lainnya (Ridwan dkk., 2024) memanfaatkan kombinasi metode SMOTE, SVM-SMOTE, KMeans-SMOTE, dan *Borderline-SMOTE*, yang diuji dengan empat algoritma klasifikasi: *Random Forest*, SVM, *Logistic Regression*, dan *Naïve Bayes*. Sementara itu, penelitian (Wicaksono dkk., 2024) menggunakan metode SMOTE, ADASYN, dan *Random Oversampling* dengan algoritma XGBoost.

Meskipun penelitian (E. Fitri dkk., 2020) menunjukkan keberhasilan dalam mengatasi data tidak seimbang (*imbalanced data*) dengan SMOTE, pendekatannya terbatas pada satu metode dan dataset yang digunakan hanya mencakup 1.629 ulasan pengguna, yang ruang lingkupnya masih bisa ditingkatkan. Penelitian (Ridwan dkk., 2024) berhasil mengimplementasikan *Borderline-SMOTE*, namun hanya memanfaatkan dataset statis dari tahun 2018 dan tidak menyertakan visualisasi data sebelum dan sesudah *oversampling* menggunakan PCA. Adapun penelitian (Wicaksono dkk., 2024), meskipun berhasil menerapkan *Random Oversampling*, hanya menggunakan dataset berisi 2.139 data dan mengandalkan satu algoritma pembelajaran mesin, yakni XGBoost.

Berdasarkan identifikasi gap masalah tersebut, penelitian ini bertujuan untuk menerapkan dan mengevaluasi beberapa metode *oversampling* yaitu SMOTE,

Borderline-SMOTE dan *Random Oversampling* guna mengatasi masalah ketidakseimbangan data (*Imbalanced* data) pada klasifikasi sentimen. Tiga algoritma pembelajaran mesin yaitu *Naïve Bayes*, *Support Vector Machine* (SVM), dan *Random Forest* digunakan untuk menguji kinerja klasifikasi sebelum dan sesudah penerapan teknik *oversampling*. Evaluasi dilakukan dengan metrik-metrik komprehensif seperti *Confusion matrix*, *Accuracy*, *Precision*, *Recall*, *F1-Score*, dan AUC-ROC. Visualisasi PCA juga digunakan untuk mempermudah pemahaman terhadap perubahan distribusi data akibat *oversampling*.

I.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang telah dipaparkan, adapun rumusan masalah pada penelitian ini adalah sebagai berikut:

1. Bagaimana performa algoritma klasifikasi *Naïve Bayes*, *Support Vector Machine* (SVM), dan *Random Forest* dalam melakukan analisis performa pada data *Imbalanced*?
2. Bagaimana pengaruh penerapan metode *oversampling*, yaitu *Random Oversampling*, SMOTE, dan *Borderline-SMOTE*, terhadap peningkatan performa model klasifikasi sentimen pada data *Imbalanced*?
3. Kombinasi metode *oversampling* dan algoritma klasifikasi mana yang memberikan hasil terbaik dalam menangani permasalahan *Imbalanced* data pada klasifikasi sentimen?

I.3 Tujuan Penelitian

Berdasarkan latar belakang masalah dan rumusan masalah, maka tujuan dari penelitian ini adalah:

1. Mengevaluasi performa algoritma klasifikasi *Naïve Bayes*, *Support Vector Machine* (SVM), dan *Random Forest* dalam menganalisis performa pada data *Imbalanced* sebelum dan sesudah penerapan metode *oversampling*.
2. Menganalisis pengaruh metode *oversampling* (*Random Oversampling*, SMOTE, dan *Borderline-SMOTE*) terhadap perbaikan performa model dalam menangani data *Imbalanced*.
3. Mengevaluasi kombinasi metode *oversampling* dan algoritma klasifikasi terbaik yang mampu meningkatkan akurasi dan kemampuan klasifikasi pada data *Imbalanced*.

I.4 Manfaat Penelitian

Adapun manfaat penelitian yang dilakukan adalah sebagai berikut:

1. Penelitian ini diharapkan dapat memberikan kontribusi pada pengembangan metode analisis sentimen, khususnya dalam mengatasi masalah *Imbalanced* data, dengan menggunakan salah satu algoritma *Naïve Bayes*, *Random Forest*, *Support Vector Machine* (SVM), menggunakan kombinasi metode *oversampling* SMOTE, Borderline SMOTE, *Random Oversampling*.
2. Hasil penelitian ini dapat digunakan sebagai referensi dalam memilih algoritma dan teknik yang tepat untuk melakukan analisis sentimen pada berbagai domain, seperti analisis sentimen terhadap produk, merek, atau isu sosial.
3. Manfaat praktis dalam meningkatkan model akurasi dengan mengatasi *Imbalanced* data, *oversampling methods* dapat memberikan prediksi yang lebih akurat, khususnya pada kelas minoritas.

I.5 Batasan Masalah

Adapun batasan masalah yang dilakukan adalah sebagai berikut:

1. Penelitian ini hanya fokus pada data teks yang bersifat opini dan mengandung sentimen, data yang digunakan hanya ulasan Tokopedia, dan menggunakan bahasa hanya Indonesia.
2. Penelitian ini hanya membandingkan tiga teknik *oversampling*, yaitu SMOTE, *Borderline-SMOTE* dan *Random Oversampling* (ROS).
3. Penelitian ini hanya membandingkan tiga klasifikasi, yaitu *Naïve Bayes*, *Support Vector Machine* (SVM), dan *Random Forest*.
4. Kinerja metode dan algoritma akan dievaluasi menggunakan metrik seperti *Accuracy*, *Precision*, *Recall*, *F1-Score*, dan AUC-ROC.