BAB I
PENDAHULUAN

1.1 Latar Belakang

Ketidakseimbangan data atau *Imbalance Data* sering menjadi masalah utama dalam *machine learning*. *Imbalance Data* terjadi ketika distribusi kelas dalam dataset tidak merata, misalnya, jumlah data sentimen *positif* jauh lebih banyak dibandingkan sentimen *negatif* atau sebaliknya. *Imbalance Data* semakin signifikan di era digital, di mana tanggapan masyarakat tersebar luas melalui media sosial, terutama X (*Twitter*). Salah satu topik yang sering menjadi perbincangan adalah kebijakan naturalisasi pemain sepak bola di Indonesia karena menimbulkan banyak perdebatan, yang memunculkan berbagai tanggapan, baik *positif*, *negatif*, maupun *netral*, sebagian pihak menilai sebagai upaya meningkatkan kualitas sepak bola di Indonesia, sementara pihak lain menghawatirkan dampaknya terhadap pemain lokal. Representasi yang tidak akurat, *overfitting*, dan bias terhadap kelas mayoritas dapat diakibatkan oleh *imbalance data*, yang merupakan penghalang lain dalam klasifikasi (Syaputra et al., 2024). *Imbalance Data* dapat menyebabkan algoritma klasifikasi lebih fokus terhadap kelas mayoritas, sehingga menurunkan kinerjanya dalam mengenali kelas minoritas yang sering kali justru menjadi informasi yang paling penting.

Beberapa metode *Imbalance Data Sampling* diterapkan untuk mengatasi *Imbalance Data*, di antaranya adalah *Random Under Sampling* (*RUS*), *Random Over Sampling* (*ROS*), dan *Synthetic Minority Oversampling Technique* (*SMOTE*). *RUS* secara acak mengurangi jumlah data dari kelas mayoritas sehingga distribusinya sama dengan kelas minoritas. Pendekatan ini sederhana namun dapat menyebabkan hilangnya informasi yang penting. Sebaliknya, *ROS* menduplikasi sampel secara acak untuk meningkatkan jumlah data kelas minoritas. Meskipun efektif dalam menyeimbangkan data, metode ini berisiko menyebabkan *overfitting*. Di sisi lain, *SMOTE* menggunakan strategi yang lebih canggih dengan menggunakan *interpolasi* untuk membuat sampel sintetis untuk kelas minoritas, yang bertujuan mengurangi risiko *overfitting* sekaligus menjaga keseimbangan data.

Penerapan metode *Imbalance Data Sampling* ini sangat berpengaruh terhadap performa algoritma klasifikasi, salah satunya adalah *Support Vector Machine* (*SVM*). Dalam analisis sentimen, *SVM* merupakan salah satu metode klasifikasi yang paling banyak digunakan. *SVM* beroperasi dengan menentukan *hyperlane* terbaik untuk memisahkan data ke dalam kelas-kelas yang berbeda. Algoritma ini berkerja dengan melakukan pembobotan melalui pembentukan pola garis yang digunakan untuk proses pembobotan dan klasifikasi (Faisal et al., 2024). Keunggulan metode ini terletak pada kemampuannya menangani data berdimensi tinggi dan memberikan hasil yang optimal pada dataset yang terstuktur dengan baik. Dibandingkan metode *machine learning* lainnya, *SVM* memiliki performa yang lebih tinggi (Oktavia et al., 2023). Performa *SVM* dapat dipengaruhi oleh ketidakseimbangan data karena model ini cenderung lebih fokus pada kelas mayoritas dan kurang fokus pada kelas minoritas, meskipun kelas minoritas memiliki peran penting dalam pendeteksian yang akurat (Islam & Agung, 2025).

Penelitian ini bertujuan untuk mengukur pengaruh dari metode *Imbalance Data Sampling* seperti *RUS*, *ROS,* dan *SMOTE* terhadap kinerja *SVM* dalam analisis sentimen. Dengan mengevaluasi performa model berdasarakan *matrix* seperti *Accuracy*, *Precision*, *Recall*, dan *F1-score*, diharapkan penelitian ini dapat memberikan rekomendasi metode yang paling efektif untuk meningkatkan kemampuan *SVM* dalam mengatasi *Imbalance Data*.

Penerapan metode sampling yang tepat sangat penting untuk meningkatkan kinerja *SVM*, khususnya dalam analisis sentimen dengan data yang tidak seimbang. Penelitian ini memberikan pemahaman yang lebih baik tentang persepsi masyarakat terhadap suatu topik tertentu, sekaligus memberikan pemahaman ilmiah tentang analisis sentimen dan *machine learning*.

1.2 Rumusan Masalah

Metode *resampling* manakah di antara *Random Under Sampling* (*RUS*), *Random Over Sampling* (*ROS*), dan *Synthetic Minority Oversampling Technique* (*SMOTE*) yang memberikan performa terbaik terhadap model *Support Vector Machine* (*SVM*) dalam menangani ketidakseimbangan data pada analisis sentimen?

1.3 Tujuan Penelitian

Menganalisis sejauh mana pengaruh metode *Random Under Sampling* (*RUS*), *Random Over Samping* (*ROS*), dan *Synthetic Minority Oversampling Technique* (*SMOTE*) dalam mengatasi ketidakseimbangan data terhadap kinerja *Support Vector Machine* (*SVM*) dalam analisis sentimen.

1.4 Batasan Penelitian

1. Data pada analisis sentimen ini berfokus terhadap topik naturalisasi pemain sepak bola di Indonesia.
2. Data yang diambil adalah dari media sosial X (*Twitter*) dengan menggunakan hashtag #Naturalisasi
3. Data yang diambil pada bulan Januari 2021 sampai Desember 2024.
4. Data yang digunakan sebanyak 1530.
5. Data yang digunakan hanya mencangkup komentar dalam bahasa Indonesia.
6. Evaluasi kinerja model dilakukan dengan menggunakan *matrix* seperti *Accuracy, Precision, Recall,* dan *F1-score* guna menentukan metode *Imbalance Data Sampling* yang terbaik.
7. Penelitian ini tidak membatasi penggunaan kata tidak baku di media sosial untuk menjaga keaslian opini pengguna.
8. Hasil dari klasifikasi berupa sentimen *positif*, *negatif,* dan *netral*.
9. Bahasa pemrograman *Python* digunakan dalam penelitian ini.

1.5 Manfaat Penelitian

1. Manfaat Teoritis

Penelitian ini berkontribusi dalam pengembangan ilmu pengetahuan, khususnya dalam *machine learning* dan analisis sentimen. Studi ini memberikan wawasan baru mengenai efektivitas penerapan metode *Imbalance Data Sampling* seperti *RUS, ROS, SMOTE,* yang dapat meningkatkan kinerja model dengan mengatasi ketidakseimbangan kelas dalam data. Penelitian ini memperluas mengenai bagaimana pendekatan metode *Imbalance Data Sampling* dapat mempengaruhi kinerja *SVM* dalam analisis sentimen.

2. Manfaat Praktis

Penelitian ini berpotensi memberikan manfaat dalam berbagai aspek, terutama dalam analisis sentimen terhadap kebijakan naturalisasi pemain sepak bola di Indonesia dengan penerapan metode *Imbalance Data Sampling* yang tepat. Penelitian ini juga dapat membantu meningkatkan *Accuracy* klasifikasi sentimen publik terhadap kebijakan tersebut. Hasil analisis yang akurat dapat memberikan wawasan kepada federasi sepak bola, pemerintah, dan pemangku kebijakan lainnya dalam menilai tanggapan masyarakat tentang kebijakan naturalisasi pemain sepak bola di Indonesia.