BAB I
PENDAHULUAN

1.1 Latar Belakang

Seiring dengan kemajuan era digital, media sosial telah bertransformasi menjadi ruang publik virtual di mana individu dapat mengartikulasikan pandangannya mengenai banyak hal, tidak terkecuali dunia olahraga. Isu yang menarik perhatian adalah naturalisasi pemain sepak bola. Opini publik sering kali dibagi 2, sentimen positif dan negatif. Untuk memahami pola opini ini, teknik analisis sentimen digunakan, khususnya metode pemrosesan bahasa alami untuk mengategorikan teks sesuai dengan polaritas emosi yang diungkapkan, yang dapat berupa sentimen positif, negatif, atau netral.

Terdapat tantangan signifikan dalam menerapkan analisis sentimen, yaitu ketidakseimbangan data. Ketidakseimbangan ini terjadi merujuk pada kondisi di mana distribusi contoh antar kelas dalam sebuah dataset tidak merata, dengan satu kelas memiliki jumlah contoh yang signifikan lebih besar dari kelas lainnya, yang dapat menyebabkan model prediksi menjadi bias ke arah kelas mayoritas (Meidianingsih et al., 2023). Dalam analisis sentimen mengenai naturalisasi pemain sepak bola, pandangan positif mungkin jauh lebih dominan daripada pandangan negatif atau netral, model *Support Vector Machine* yang digunakan untuk klasifikasi mungkin mengalami kesulitan mendeteksi contoh dari kelas minoritas (Tiara et al., 2024).

Ketidakseimbangan data yang tinggi mengakibatkan model pembelajaran mesin gagal memprediksi proporsi kelas yang kecil dengan tepat (Lee et al., 2020). Hal ini berdampak negatif pada kinerja model, terutama dalam hal akurasi klasifikasi dan kemampuan untuk menggeneralisasi ke data baru. Maka dibutuhkan metode penyelesaian masalah dan memastikan hasil analisis lebih representatif dan akurat.

Solusi menanggulangi ketidakseimbangan data yakni teknik *imbalance data sampling*, yang bertujuan untuk menyeimbangkan distribusi kelas dalam kumpulan data sebelum melakukan proses pelatihan model. Dalam penelitian ini, *Borderline-*

SVM, SVM-SMOTE, dan ADASYN digunakan sebagai metode pengambilan sampel. *Borderline-SMOTE* merupakan teknik *oversampling* yang menargetkan sampel minoritas yang dekat dengan batas keputusan, sehingga meningkatkan kemampuan pembelajaran model pada data yang sulit diklasifikasikan (Lee et al., 2020). SVM-SMOTE adalah metode yang menggabungkan *Support Vector Machine* dan *Synthetic Minority Over-sampling Technique* guna menghasilkan contoh sintetis baru berdasarkan *support vectors*, yang membantu dalam menciptakan representasi kelas minoritas yang lebih efisien (Meidianingsih et al., 2023). ADASYN adalah metode adaptif yang secara otomatis menyesuaikan jumlah sampel sintetis yang dihasilkan berdasarkan distribusi data minoritas, membantu model untuk lebih fokus pada sampel yang lebih sulit diklasifikasikan (Tiara et al., 2024).

Dalam konteks analisis sentimen naturalisasi pemain sepak bola menggunakan data yang diperoleh dari X, penerapan teknik *Borderline-SMOTE*, SVM-SMOTE, dan ADASYN diharapkan dapat meningkatkan kinerja prediksi model *Support Vector Machine*. Distribusi data yang lebih seimbang memungkinkan model mempelajari pola sentimen yang lebih menyeluruh, menghasilkan klasifikasi yang lebih akurat dan lebih sedikit bias terhadap kelas mayoritas.

Lebih lanjut, studi ini berupaya untuk memberikan sumbangsih terhadap peningkatan efektivitas metode klasifikasi teks yang ada, terutama mengingat masalah ketidakseimbangan data di media sosial. Dengan menganalisis data dari berbagai sumber, termasuk X (*Twitter*), dapat menggali lebih dalam tentang opini publik terhadap kebijakan ini.

Penelitian ini bertujuan mengevaluasi efektivitas metode *imbalanced data sampling*, yaitu *Borderline-SMOTE*, SVM-SMOTE, dan ADASYN, dalam meningkatkan kinerja *Support Vector Machine (SVM)* untuk menganalisis opini publik mengenai kebijakan naturalisasi pemain sepak bola di Indonesia, merekomendasi metode *sampling* yang paling efektif berdasarkan hasil evaluasi akurasi, presisi, *recall*, dan *F1-score* dari model yang diterapkan pada dataset tidak seimbang. Hasil penelitian ini diharapkan dapat memberikan wawasan dan

rekomendasi tentang metode *sampling* yang paling tepat untuk mengatasi ketidakseimbangan data dalam analisis sentimen media sosial.

1.2 Rumusan Masalah

Metode *imbalance data sampling* manakah di antara *Borderline-SMOTE*, SVM-SMOTE, dan ADASYN yang paling efektif dalam meningkatkan *accuracy*, *precision*, *recall*, dan *F1-score* model SVM untuk menganalisis opini publik mengenai kebijakan naturalisasi pemain sepak bola di Indonesia?

1.3 Tujuan Penelitian

Mengevaluasi efektivitas tiga metode *Imbalance Data Sampling*, yaitu *Borderline-SMOTE*, SVM-SMOTE, dan ADASYN, dalam meningkatkan *accuracy*, *precision*, *recall*, dan *F1-score* model *Support Vector Machine (SVM)* untuk menganalisis opini publik terhadap kebijakan naturalisasi pemain sepak bola di Indonesia.

1.4 Manfaat Penelitian

1. Manfaat Teoritis

Berkontribusi pada pengembangan teori pembelajaran mesin, khususnya dengan menerapkan metode *imbalance data sampling* pada SVM untuk analisis sentimen. Hasil penelitian ini tidak hanya memperdalam pemahaman tentang cara mengenai masalah *class imbalance*, tetapi juga dapat memberikan wawasan baru tentang efektivitas berbagai metode *sampling* dalam analisis sentiment, dijadikan referensi untuk studi berikutnya di bidang klasifikasi teks dan analisis sentimen.

2. Manfaat Praktis

Berkontribusi bagi praktisi data *science* dan pengembang sistem dalam memilih metode *imbalance data sampling* yang efektif untuk data yang tidak seimbang. Organisasi seperti PSSI dan lembaga pemerintah lainnya dapat menggunakan hasil penelitian ini untuk menganalisis opini publik terhadap kebijakan. Hal ini memungkinkan mereka membuat keputusan yang lebih tepat, seperti mengembangkan strategi komunikasi yang lebih baik atau mengadaptasi kebijakan berdasarkan opini publik.

1.5 Batasan Penelitian

1. Fokus penelitian ini hanya pada tiga metode *imbalance data sampling*, yaitu *Borderline-SMOTE*, *SVM-SMOTE,* dan *ADASYN.*
2. Penelitian ini menggunakan 1530 data *tweet* di X (*Twitter)* dengan *hashtag* #Naturalisasi, yang dikumpulkan dari bulan Januari 2021 hingga Desember 2024.
3. Kinerja model dievaluasi berdasarkan *matrix* seperti *accuracy*, *precision*, *recall*, dan *F1-score* guna membandingkan efektivitas masing-masing metode *imbalance data sampling.*
4. Penelitian ini tidak membatasi penggunaan kata tidak baku di media sosial untuk menjaga keaslian opini pengguna.
5. Data yang digunakan hanya mencangkup komentar dalam bahasa Indonesia.
6. Hasil dari klasifikasi berupa sentimen positif, negatif, dan netral.
7. Penelitian ini menggunakan Bahasa pemrograman *Python*