BAB I
PENDAHULUAN

1.1 Latar Belakang

Naturalisasi adalah proses mendapatkan status kewarganegaraan bagi warga asing yang memenuhi syarat dan prosedur dalam naturalisasi. Di Indonesia, kebijakan naturalisasi memiliki beberapa syarat yang harus dipenuhi oleh seseorang yang ingin mendapatkan status tersebut. Syarat yang sering digunakan di naturalisasi timnas indonesia ini didasarkan pada keturunan orang tua yang biasanya juga disebut sebagai pemain diaspora [1]. Dalam dunia sepak bola, naturalisasi sudah banyak dilakukan di berbagai negara di seluruh benua. Naturalisasi dilakukan oleh suatu negara untuk mendapatkan pemain berpotensi yang dapat meningkatkan kekuatan tim nasional. Pada era Shin Tae-yong, PSSI banyak melakukan naturalisasi pemain keturunan untuk meningkatkan prestasi timnas Indonesia. Proses ini dilakukan dengan alasan jelas, terutama mengingat timnas indonesia memiliki target untuk lolos ke Piala Dunia 2026. Topik naturalisasi ini menimbulkan banyak pro dan kontra di masyarakat dan menjadi salah satu topik yang sering dibahas di media sosial X [2].

Media sosial Twitter, yang sangat banyak digunakan oleh masyarakat Indonesia, menjadi platform utama untuk mendiskusikan topik ini. Twitter tidak hanya digunakan untuk berkomunikasi, tetapi juga untuk menyampaikan pendapat melalui komentar. Dalam topik naturalisasi, terdapat komentar yang mendukung Langkah naturalisasi ini, mereka berpendapat bahwa pemain naturalisasi dapat meningkatkan kekuatan timnas Indonesia berdasarkan pengalaman mereka bermain di luar negeri. Namun, banyak juga komentar yang menyuarakan keresahan atau kritik bahwa naturalisasi pemain akan mengurangi kesempatan bagi pemain lokal untuk berkembang dan tampil di pertandingan internasional [3].

Untuk memahami pendapat masyarakat tentang naturalisasi, diperlukan metode analisis sentimen. Analisis sentimen adalah proses memahami dan mengolah data teks untuk mendapatkan informasi dan mendeteksi opini terhadap suatu objek atau kasus [4]. Analisis sentimen digunakan untuk mengklasifikasikan

komentar pengguna X ke dalam beberapa kategori, di mana penelitian ini terdapat tiga kategori sentimen: positif, netral, dan negatif [5].

Penelitian analisis sentimen sudah banyak digunakan untuk menlihat opini masyarakat terhadap suatu kasus. Banyak metode yang bisa digunakan untuk melatih model seperti Naïve bayes, Decision Tree, SVM, LSTM, dan BERT. Metode yang akan digunakan adalah menggunakan metode berbasis deep learning yaitu BERT. BERT (Bidirectional Encoder Representations from Transformers) adalah arsitektur transformer yang sudah dilatih menggunakan wikipedia Bahasa inggris 2500 juta kata dan BookCorpus dengan 800 juta. BERT dapat mencapai performa yang baik walau dengan menggunakan dataset kecil hanya perlu melakukan fine-tuning [6]. BERT memiliki enam lapisan Transformer pada setiap encoder dan decoder, yang membuat proses pelatihan menjadi rumit, dengan konfigurasi kompleks, waktu pelatihan yang lama, dan biaya tinggi. Namun, model pra-latih BERT dari Google tersedia sebagai open-source, sehingga dapat digunakan tanpa harus membangun model dari awal. Pemrosesan BERT dimulai dari kata dan representasi embedding, dengan setiap lapisan menggunakan multi-head attention untuk menghasilkan representasi perantara. Setiap token akan memiliki 12 representasi perantara di 12 lapisan model BERT [7].

Adapun model BERT yang sudah dilatih secara khusus untuk bahasa Indonesia yang disebut IndoBert. IndoBert adalah bentuk implementasi dari BERT yang sudah dilatih secara khusus untuk memahami data teks berbahasa Indonesia, yang memungkinkan model bisa menangkap konteks Bahasa Indonesia yang lebih baik [8]. IndoBert memiliki korpus lebih dari 220 juta kata bahasa Indonesia dan sebagian besar berasal dari Wikipedia Bahasa Indonesia yang berjumlah 75 kata, Indonesia Web Corpus 90 juta kata dan Kompas, Tempo, Artikel Liputan6 55 juta kata [9].

Untuk mengevaluasi kinerja IndoBERT, penelitian ini juga menggunakan metode lain yang berfokus pada perbandingan data balancing guna mengatasi ketidakseimbangan data. Metode pertama yang digunakan adalah wordnet yaitu sionim , random insertion, dan random deletion, kemudian metode yang kedua adalah SMOTE (Synthetic Minority Oversampling Technique). Augmentasi sinonim merupakan proses meningkatkan variasi data teks dengan mengganti kata

dengan sinonimnya. Tujuan nya yaitu memperbanyak data teks tanpa mengubah makna teks itu sendiri. Metode sinonim ini menggunakan WordNet untuk mengelompokan kata ke sinonimnya, yang bisa memungkinkan mengganti kata sesuai konteks kalimat yang ada [10]. Random deletion adalah menghapus secara acak setiap kata dalam sebuah kalimat [11]. Sedangkan SMOTE adalah metode resampling yang digunakan untuk menyamakan kelas minoritas dengan cara menyisipkan sampel sintesis pada kelas kelas minoritas [12].

Beberapa penelitian terbaru menunjukkan keunggulan algoritma Naïve Bayes dalam analisis sentimen. Dava Rizky Perwira Jaya dan Sri Lestari (2024) menemukan bahwa Naïve Bayes mencapai akurasi 85,4% dalam analisis sentimen terhadap tim Indonesia U-23, lebih baik dibandingkan dengan KNN yang hanya memperoleh 71,8% [13]. Hal serupa juga dibuktikan oleh Billy Franko, Nicholas Wilyanto, dan Hafiz Irsyad (2024) dalam penelitian mereka tentang naturalisasi pemain di YouTube, di mana Naïve Bayes juga meraih akurasi 85,4%, sementara Decision Tree hanya mencapai 71,8% [14]. Namun, penelitian oleh Michelle Graciela, Rikky, dan Hafiz Irsyad (2024) menemukan bahwa model KNN tanpa SMOTE dapat mencapai akurasi hampir sempurna sebesar 98%, tetapi penerapan SMOTE justru menurunkan performa model pada metrik evaluasi [15]. Selain itu, penelitian oleh Leno Dwi Cahya dan tim (2023) berfokus pada penanganan ketidakseimbangan dalam dataset multi-label untuk analisis sentimen dan emosi dalam bahasa Indonesia. Mereka menggunakan teknik SMOTE untuk menghasilkan data sintetis bagi kelas minoritas dan augmentasi data untuk meningkatkan variasi teks. Hasilnya menunjukkan bahwa SMOTE dapat meningkatkan akurasi hingga 82%, sementara augmentasi data memberikan peningkatan akurasi hingga 78% [16].

Penelitian ini dilakukan dengan tujuan untuk mengevaluasi efektivitas dua teknik penyeimbangan data, yaitu WordNet dan Synthetic Minority Oversampling Technique (SMOTE), dengan menggunakan model IndoBERT dalam analisis sentimen terhadap kebijakan naturalisasi pemain Timnas Indonesia di media sosial X. Model IndoBERT dipilih karena merupakan salah satu model bahasa berbasis BERT yang dikembangkan secara khusus untuk memahami konteks bahasa Indonesia secara mendalam. Kemampuannya dalam mengenali makna kata,

struktur kalimat, serta nuansa emosional dalam opini masyarakat menjadikannya relevan untuk digunakan dalam menganalisis sentimen publik terhadap kebijakan nasional yang berkaitan dengan isu sosial dan identitas bangsa, seperti kebijakan naturalisasi pemain sepak bola.

Dua teknik penyeimbangan data, yakni WordNet dan SMOTE, digunakan karena keduanya telah banyak diterapkan dalam berbagai penelitian terdahulu dengan hasil yang bervariasi tergantung pada karakteristik data. Beberapa studi menunjukkan bahwa SMOTE efektif dalam meningkatkan akurasi pada data tidak seimbang, sementara WordNet berperan dalam memperkaya makna semantik dari data teks. Oleh karena itu, penelitian ini membandingkan efektivitas kedua metode tersebut dalam konteks analisis sentimen berbahasa Indonesia. Topik kebijakan naturalisasi Timnas Indonesia dipilih karena isu ini tengah ramai diperbincangkan di media sosial dan menimbulkan beragam opini masyarakat. Dengan demikian, penelitian ini tidak hanya menguji performa model dan teknik penyeimbangan data, tetapi juga memberikan gambaran mengenai bagaimana masyarakat menanggapi kebijakan naturalisasi tersebut sebagai bagian dari dinamika sosial dalam dunia olahraga Indonesia.

Data dikumpulkan melalui proses web crawling di platform Twitter dalam dua tahap pengambilan. Tahap pertama dilakukan pada periode 28 September 2023 hingga 28 September 2024, namun sempat terhenti di tengah proses karena kendala teknis, sehingga dilakukan tahap kedua pada periode 2 November 2023 hingga 2 November 2024 untuk menambah jumlah data yang diperoleh. Rentang waktu tersebut menunjukkan periode unggahan tweet yang menjadi sumber data penelitian, dengan fokus pada tweet yang dipublikasikan pada tanggal-tanggal tersebut. Data yang diperoleh tidak mencakup seluruh tweet dalam periode itu secara menyeluruh, melainkan terambil secara acak berdasarkan hasil crawling yang berhasil diperoleh dari sistem. Secara keseluruhan, data yang berhasil dikumpulkan dari dua kali proses crawling berjumlah 7.866 tweet, yang digunakan sebagai dasar analisis sentimen terhadap kebijakan naturalisasi Timnas Indonesia dengan kata kunci ‘naturalisasi timnas indonesia’.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah tersebut, maka peneliti mengambil rumusan masalah sebagai berikut:

a. Bagaimana sentimen komentar masyarakat di media sosial X terkait proses naturalisasi timnas Indonesia?
b. Apa perbandingan efektivitas teknik data balancing WordNet dan SMOTE dalam meningkatkan akurasi model IndoBERT untuk analisis sentimen?
c. Apa pengaruh dari penggunaan teknik data balancing terhadap performa model IndoBERT dalam mengklasifikasikan komentar menjadi kategori positif, negatif, dan netral?

1.3 Tujuan Penelitian

Berdasarkan dari rumusan masalah tersebut, maka tujuan dari penelitian ini sebagai berikut:

a. Menganalisis sentimen komentar masyarakat di media sosial X mengenai naturalisasi timnas Indonesia, untuk mengetahui opini yang bersifat positif, negatif, atau netral.
b. Mengevaluasi efektivitas teknik data balancing teks WordNet dan SMOTE dalam meningkatkan akurasi model IndoBERT dalam analisis sentimen.
c. Membandingkan performa model IndoBERT yang dilatih dengan teknik augmentasi WordNet dan SMOTE dalam klasifikasi komentar masyarakat terkait naturalisasi.

1.4 Batasan Masalah

Pada batasan masalah ini, peneliti membatasi beberapa permasalahan yang perlu agar pembahasan pada penelitian ini dapat tertuju dengan baik, yaitu:

a. Penelitian ini berfokus pada komentar masyarakat di media sosial X terkait kebijakan naturalisasi Timnas Indonesia, dengan data yang dikumpulkan melalui dua kali proses crawling pada tanggal 28 September 2024 dan 2 November 2024. Setiap proses mencakup rentang waktu satu tahun unggahan, namun data yang terambil bersifat acak dan tidak mencakup seluruh tweet dalam periode tersebut, meskipun rentang waktunya telah ditetapkan.

b. Data yang dianalisis terbatas pada 7.866 tweet yang diambil menggunakan kata kunci 'naturalisasi timnas indonesia'.
c. Model yang digunakan dalam penelitian ini adalah IndoBERT, dengan dua teknik data balancing yang dibandingkan: WordNet dan SMOTE.
d. Klasifikasi sentimen dibatasi pada tiga kategori: positif, negatif, dan netral.
e. Menggunakan bahasa pemrograman Python.
f. Menggunakan Google Collaboratory.