BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi pemrosesan bahasa alami atau NLP (*Natural Language Processing*), telah membawa transformasi signifikan dalam cara manusia berinteraksi dengan data tekstual. Salah satu cabang NLP yang paling banyak diteliti adalah analisis sentimen, yaitu proses komputasi untuk mengidentifikasi dan mengklasifikasikan opini atau emosi yang terkandung dalam suatu teks (Liu, 2012). Analisis sentimen memiliki peran strategis dalam berbagai domain, mulai dari pemantauan opini publik di media sosial, evaluasi produk, hingga pengambilan keputusan berbasis data di sektor bisnis dan pemerintahan (Birjali dkk., 2021). Seiring berkembangnya era digital, volume data tekstual yang dihasilkan dari *platform* daring meningkat secara eksponensial, sehingga kebutuhan akan metode analisis sentimen yang akurat dan efisien menjadi semakin mendesak.

Revolusi dalam bidang NLP dimulai dengan diperkenalkannya arsitektur Transformer oleh Vaswani dkk. (2017) yang mengusung mekanisme *self-attention* untuk memproses sekuens teks secara paralel. Arsitektur ini kemudian menjadi fondasi bagi model *Bidirectional Encoder Representations from Transformers* (BERT) yang dikembangkan oleh Devlin dkk. (2019). BERT mampu mempelajari representasi kontekstual kata secara bidireksional, sehingga mengungguli pendekatan-pendekatan sebelumnya pada berbagai *benchmark* NLP. Keberhasilan

BERT mendorong pengembangan varian-varian model pra latih (*pre-trained)* yang diadaptasi untuk bahasa dan konteks linguistik tertentu, termasuk model *multilingual* seperti Multilingual BERT (mBERT), XLM-RoBERTa (Conneau dkk., 2020), serta model monolingual seperti IndoBERT (Koto dkk., 2020; Wilie dkk., 2020).

Meskipun kemajuan tersebut sangat pesat, sebagian besar penelitian NLP masih berfokus pada bahasa-bahasa dengan sumber daya tinggi (*high-resource languages*) seperti bahasa Inggris, Mandarin, dan Spanyol. Indonesia, sebagai negara dengan lebih dari 700 bahasa daerah (Eberhard dkk., 2025), menghadapi tantangan serius dalam pengembangan teknologi NLP untuk bahasa-bahasa lokalnya. Aji dkk. (2022) mengidentifikasi bahwa mayoritas bahasa daerah di Indonesia termasuk dalam kategori bahasa dengan sumber daya terbatas (*low-resource language)* karena minimnya ketersediaan korpora digital, alat bantu linguistik, maupun model bahasa yang secara khusus dilatih untuk bahasa-bahasa tersebut. Kondisi ini menciptakan kesenjangan teknologi yang signifikan antara bahasa Indonesia sebagai bahasa nasional dan ratusan bahasa daerah lainnya.

Bahasa Bali merupakan salah satu bahasa daerah di Indonesia yang termasuk dalam rumpun bahasa Austronesia, subkelompok Bali–Sasak–Sumbawa (Arka, 2003). Berdasarkan data sensus nasional tahun 2000, bahasa Bali dituturkan oleh sekitar 3,3 juta orang, namun Badan Kebudayaan Bali memperkirakan pada tahun 2011 bahwa jumlah penutur aktif yang masih menggunakan bahasa Bali dalam kehidupan sehari-hari tidak lebih dari satu juta orang. Ethnologue mengklasifikasikan bahasa Bali sebagai bahasa yang terancam punah, di mana

anak-anak tidak lagi secara lazim mempelajari dan menggunakan bahasa ini (Eberhard dkk., 2025). Pergeseran bahasa ini, yang disebabkan oleh dominasi bahasa Indonesia dan bahasa Inggris di ranah pendidikan serta aktivitas ekonomi, menjadikan upaya pelestarian bahasa Bali melalui teknologi digital semakin penting dan mendesak. Pergeseran ini juga terkonfirmasi pada tataran lokal. Sosiawan dkk. (2021) menemukan bahwa pada keluarga muda di Kota Singaraja, penggunaan bahasa Bali oleh anak-anak sudah sangat terbatas dan cenderung tergantikan oleh bahasa Indonesia, menandakan kemunduran transmisi bahasa Bali antargenerasi.

Salah satu upaya penting dalam membangun sumber daya NLP untuk bahasa-bahasa daerah Indonesia adalah proyek NusaX yang dikembangkan oleh Winata dkk. (2023). NusaX merupakan korpus paralel multibahasa berkualitas tinggi yang mencakup 12 bahasa, yaitu bahasa Indonesia, bahasa Inggris, dan 10 bahasa daerah Indonesia, termasuk bahasa Bali. *Dataset* ini menyediakan data anotasi sentimen tiga kelas (positif, netral, negatif) yang diterjemahkan oleh penutur asli, sehingga tergolong sumber data beranotasi berkualitas tinggi untuk penelitian analisis sentimen pada bahasa-bahasa *low-resource* Indonesia. NusaX juga menyediakan leksikon bilingual yang dapat dimanfaatkan untuk keperluan audit linguistik terhadap model-model bahasa.

Dalam konteks analisis sentimen berbahasa Bali, terdapat tiga model *pre-trained* berbasis Transformer yang relevan untuk dievaluasi. Pertama, IndoBERT (Wilie dkk., 2020) yang dilatih secara khusus menggunakan korpus bahasa Indonesia berukuran lebih dari empat miliar kata. Mengingat bahasa Indonesia dan

bahasa Bali sama-sama tergolong rumpun Austronesia dan memiliki sejumlah irisan kosakata, diduga *tokenizer* IndoBERT memiliki cakupan yang lebih relevan terhadap kosakata bahasa Bali dibandingkan model *multilingual*. Dugaan ini menjadi salah satu hal yang diuji secara empiris dalam penelitian ini melalui audit *tokenizer*. Kedua, Multilingual BERT atau mBERT (Devlin dkk., 2019) yang dilatih pada korpus Wikipedia dari 104 bahasa menggunakan *tokenizer WordPiece*. Ketiga, XLM-RoBERTa (Conneau dkk., 2020) yang dilatih pada data CommonCrawl dari 100 bahasa dengan ukuran lebih dari dua *terabyte* menggunakan *tokenizer* SentencePiece. Masing-masing model memiliki strategi prapelatihan dan mekanisme tokenisasi yang berbeda, sehingga berpotensi menghasilkan representasi yang berbeda pula terhadap teks berbahasa Bali.

Salah satu aspek kritis yang belum banyak diteliti adalah bagaimana *tokenizer* dari masing-masing model menangani kosakata bahasa Bali. Rust dkk. (2021) menunjukkan bahwa kualitas *tokenizer* berpengaruh signifikan terhadap performa model *multilingual* pada bahasa-bahasa *low-resource*. Ketika *tokenizer* memecah satu kata menjadi terlalu banyak *subword* (fragmentasi tinggi), representasi semantik kata tersebut menjadi terdilusi, yang berpotensi menurunkan akurasi klasifikasi. Metrik seperti *fertility rate* (rasio token per kata), *perfect match rate* (persentase kata yang dikenali utuh), dan *Out-of-Vocabulary* (OOV) *rate* menjadi indikator penting untuk mengevaluasi sejauh mana suatu *tokenizer* mampu mengenali dan merepresentasikan kosakata bahasa target secara memadai (Petrov dkk., 2023).

Di sisi lain, model-model berbasis Transformer dikenal sebagai *black-box*

*mode*l karena kompleksitas arsitekturnya yang menyulitkan interpretasi proses pengambilan keputusan. Untuk mengatasi keterbatasan ini, pendekatan *Explainable Artificial Intelligence* (XAI) menjadi semakin penting dalam penelitian NLP (Diwali dkk., 2023). Salah satu metode XAI yang banyak digunakan adalah SHAP (*SHapley Additive exPlanations*), yang dikembangkan oleh Lundberg dan Lee (2017) berdasarkan teori nilai Shapley dari teori permainan kooperatif. SHAP mampu menghitung kontribusi setiap fitur (dalam konteks NLP, setiap token) terhadap prediksi model secara konsisten dan adil. Mosca dkk. (2022) menunjukkan bahwa SHAP dapat diintegrasikan secara efektif dengan model-model berbasis Transformer untuk menghasilkan penjelasan yang transparan dan dapat dipertanggungjawabkan.

Beberapa penelitian terdahulu telah mengeksplorasi analisis sentimen pada bahasa daerah Indonesia maupun penerapan model *pre-trained* berbasis Transformer untuk bahasa Indonesia, namun masing-masing masih memiliki keterbatasan. Wibawa dan Pramartha (2023) melakukan analisis sentimen pada teks berbahasa Bali menggunakan Multinomial *Naïve Bayes* dengan TF-IDF dan *Bag of Words*, mencapai akurasi 91,5%, namun pendekatan yang digunakan masih terbatas pada metode machine learning konvensional tanpa memanfaatkan model *pre-trained* berbasis Transformer. Studi perbandingan algoritma konvensional juga banyak dilakukan, seperti Ariyani dkk. (2025) yang membandingkan *Naïve Bayes* dan *K-Nearest Neighbor* dengan pembobotan TF-IDF pada sentimen opini publik mengenai virus Corona di Twitter, dengan *Naïve Bayes* (akurasi 0,83) lebih unggul dibandingkan *K-Nearest Neighbor* (0,78). Pada domain yang berbeda, Saputri dkk.

(2024) membandingkan *Naïve Bayes* dan *Long Short-Term Memory* (LSTM) untuk sentimen komentar mahasiswa terhadap pelayanan daring, dengan *Naïve Bayes* (akurasi 83,69%) justru mengungguli LSTM (53,12%) akibat keterbatasan distribusi data pelatihan. Pendekatan deep learning kemudian dikembangkan lebih jauh oleh Santhi dkk. (2026) yang menerapkan LSTM dengan TF-IDF dan penyeimbangan kelas SMOTE untuk klasifikasi sentimen tiga kelas terhadap isu utang negara, mencapai akurasi 85%. Seiring perkembangan model *pre-trained*, Prabowo dan Sanjaya (2024) menerapkan *transfer learning* IndoBERT untuk analisis sentimen bahasa Jawa Ngoko Lugu dengan akurasi 90%, menunjukkan potensi *transfer learning* lintas bahasa serumpun, namun hanya menguji satu model tanpa perbandingan. Pada domain bahasa Indonesia, Tarwoto (2025) menggunakan IndoBERT untuk analisis sentimen ulasan aplikasi Mobile JKN dengan akurasi 97,28%, sementara Alvin dan Winarsih (2025) membandingkan IndoBERT dan IndoBERTweet dengan algoritma klasik (SVM, *Logistic Regression, Random Forest*), menunjukkan keunggulan model Transformer dengan F1-Score 0,93. Penelitian-penelitian tersebut, baik yang menggunakan metode konvensional maupun model berbasis Transformer, umumnya mengevaluasi model hanya dari sisi performa klasifikasi tanpa menelusuri faktor yang mendasari perbedaan performa maupun dasar pengambilan keputusan model.

Meskipun pendekatan *machine learning* konvensional dapat mencapai akurasi yang tinggi pada tugas klasifikasi sentimen, pendekatan tersebut memiliki keterbatasan yang relevan dengan fokus penelitian ini. Metode seperti Multinomial Naïve Bayes dengan TF-IDF dan *Bag of Words* merepresentasikan teks

berdasarkan frekuensi kemunculan kata tanpa mempertimbangkan urutan dan konteks kemunculannya (Jurafsky & Martin, 2023; Liu, 2012), sehingga kurang mampu menangkap makna yang bergantung pada konteks. Sebaliknya, arsitektur Transformer dengan mekanisme *self-attention* (Vaswani dkk., 2017) menghasilkan representasi kata yang bersifat kontekstual, di mana kata yang sama dapat direpresentasikan secara berbeda bergantung pada konteks kalimatnya (Devlin dkk., 2019). Selain itu, pendekatan konvensional tidak melibatkan mekanisme tokenisasi *subword* sebagaimana yang digunakan pada model *pre-trained* (Kudo & Richardson, 2018; Sennrich dkk., 2016), padahal kualitas tokenisasi telah ditunjukkan berpengaruh terhadap performa model pada bahasa *low-resource* (Rust dkk., 2021). Hal ini menjadikan analisis mengenai representasi kosakata bahasa Bali pada level tokenisasi, yang merupakan salah satu fokus utama penelitian ini, tidak dapat dilakukan menggunakan pendekatan konvensional. Lebih lanjut, penelitian ini bertujuan menganalisis dasar pengambilan keputusan model menggunakan SHAP (Lundberg & Lee, 2017), yang penerapannya pada model berbasis *Transformer* telah dikembangkan untuk menghasilkan atribusi pada level token (Kokalj dkk., 2021; Mosca dkk., 2022). Oleh karena itu, penggunaan model *pre-trained* berbasis *Transformer* dalam penelitian ini tidak semata-mata ditujukan untuk mengejar akurasi tertinggi, melainkan untuk memungkinkan evaluasi multi-dimensi yang mencakup performa, kualitas tokenisasi, dan interpretabilitas. Selain itu, perbandingan nilai akurasi antar penelitian perlu mempertimbangkan perbedaan *dataset*, jumlah kelas sentimen, dan metrik evaluasi yang digunakan, sehingga akurasi yang lebih tinggi pada suatu penelitian tidak

serta-merta menunjukkan superioritas metode secara mutlak.

Berdasarkan tinjauan literatur dan penelitian terdahulu di atas, teridentifikasi celah penelitian yang signifikan: belum ditemukan studi yang secara komprehensif membandingkan performa IndoBERT, mBERT, dan XLM-RoBERTa terhadap analisis sentimen teks berbahasa Bali, sekaligus menganalisis hubungan antara kualitas tokenisasi dengan performa klasifikasi, dan menginterpretasikan proses pengambilan keputusan model menggunakan SHAP. Penelitian-penelitian sebelumnya yang menggunakan dataset NusaX umumnya melaporkan performa secara agregat tanpa mendalami aspek tokenisasi dan interpretabilitas untuk bahasa Bali secara spesifik. Oleh karena itu, penelitian ini bertujuan untuk mengisi celah tersebut melalui pendekatan evaluasi multi-dimensi yang mengintegrasikan analisis performa klasifikasi, audit *tokenizer* berbasis leksikon bahasa Bali, dan interpretasi model menggunakan SHAP.

Berdasarkan celah penelitian tersebut, penelitian ini memberikan beberapa kontribusi ilmiah. Pertama, secara metodologis, penelitian ini menghasilkan kerangka evaluasi multi-dimensi yang mengintegrasikan tiga aspek analisis, yaitu performa klasifikasi, kualitas tokenisasi, dan interpretabilitas model, dalam satu prosedur evaluasi yang terpadu. Kerangka ini berpotensi direplikasi untuk penelitian pada bahasa daerah lainnya dengan memperhatikan karakteristik dataset yang digunakan, termasuk kemungkinan percampuran dengan bahasa Indonesia. Kedua, penelitian ini merupakan studi komparatif yang secara sistematis mengevaluasi tiga arsitektur tokenisasi berbeda, yaitu monolingual WordPiece (IndoBERT), multilingual WordPiece (mBERT), dan multilingual SentencePiece

(XLM-RoBERTa), terhadap kosakata bahasa Bali melalui pendekatan audit tokenizer berbasis leksikon dengan *vocabulary lookup* yang konsisten. Ketiga, penelitian ini menyediakan bukti empiris mengenai keterkaitan antara kualitas tokenisasi, performa klasifikasi, dan pola interpretabilitas model pada bahasa *low-resource*, sekaligus menunjukkan bahwa kesamaan kosakata antara data prapelatihan dan *dataset* tampak lebih berpengaruh terhadap performa dibandingkan skala model. Kontribusi-kontribusi ini diharapkan dapat menjadi landasan bagi pengembangan penelitian pemrosesan bahasa alami untuk bahasa-bahasa daerah Indonesia yang berstatus *low-resource*.

1.2 Identifikasi Masalah

Berdasarkan latar belakang yang telah diuraikan, dapat diidentifikasi beberapa masalah sebagai berikut:

1) Bahasa Bali termasuk dalam kategori *low-resource language* yang memiliki keterbatasan sumber daya NLP, sehingga model-model *pre-trained* yang tersedia belum tentu mampu merepresentasikan kosakata dan struktur bahasa Bali secara optimal.
2) *Tokenizer* dari model-model *pre-trained multilingual* dan monolingual Indonesia memiliki mekanisme tokenisasi yang berbeda (WordPiece pada mBERT dan IndoBERT, SentencePiece pada XLM-RoBERTa), yang berpotensi menghasilkan *fertility rate, perfect match rate,* dan *OOV rate* yang berbeda terhadap kosakata bahasa Bali.

3) Belum adanya kajian yang menghubungkan secara empiris antara kualitas tokenisasi (*fertility rate, perfect match rate, OOV rate*) dengan performa klasifikasi sentimen pada teks berbahasa Bali.
4) Model-model berbasis Transformer bersifat *black-box*, sehingga sulit untuk memahami mengapa suatu model memberikan prediksi tertentu, khususnya pada teks dalam bahasa yang bukan merupakan bahasa utama pelatihan model.
5) Belum tersedia studi komparatif yang mengintegrasikan evaluasi performa, audit *tokenizer*, dan interpretasi XAI secara komprehensif untuk analisis sentimen bahasa Bali.

1.3 Pembatasan Masalah

Agar penelitian ini tetap fokus dan terukur, maka ditetapkan batasan-batasan sebagai berikut:

1) Model *pre-trained* yang dievaluasi terbatas pada tiga model, yaitu IndoBERT (indobenchmark/indobert-base-p1), Multilingual BERT (bert-base-multilingual-uncased), dan XLM-RoBERTa (xlm-roberta-base).
2) Audit *tokenizer* dilakukan terhadap kosakata bahasa Bali yang terdapat dalam leksikon NusaX, bukan terhadap seluruh kosakata bahasa Bali yang ada.
3) Metode XAI yang digunakan terbatas pada SHAP (*SHapley Additive exPlanations*) dengan pendekatan *partition explainer*.

1.4 Rumusan Masalah

Berdasarkan identifikasi masalah dan batasan penelitian di atas, rumusan masalah dalam penelitian ini adalah sebagai berikut:

1) Bagaimana prosedur analisis sentimen menggunakan IndoBERT, mBERT, dan XLM-RoBERTa terhadap teks berbahasa Bali dari dataset NusaX?
2) Bagaimana perbandingan performa model IndoBERT, mBERT, dan XLM-RoBERTa dalam melakukan klasifikasi sentimen pada teks berbahasa Bali dari dataset NusaX?
3) Bagaimana interpretasi proses pengambilan keputusan masing-masing model menggunakan SHAP dalam mengklasifikasikan sentimen teks berbahasa Bali?

1.5 Tujuan Penelitian

Sejalan dengan rumusan masalah, tujuan penelitian ini adalah:

1) Merancang dan mengimplementasikan prosedur analisis sentimen teks berbahasa Bali menggunakan tiga model pre-trained berbasis Transformer (IndoBERT, mBERT dan XLM-RoBERTa) pada dataset NusaX.
2) Mengevaluasi dan membandingkan performa klasifikasi sentimen ketiga model melalui metrik akurasi dan F1-Score Macro, uji signifikansi McNemar's test, serta audit tokenizer berbasis leksikon untuk menganalisis hubungan antara kualitas tokenisasi dengan performa klasifikasi.

3) Menginterpretasi proses pengambilan keputusan masing-masing model menggunakan metode SHAP (SHapley Additive Explanations) serta mengintegrasikan temuan dari dimensi performa, tokenisasi, dan interpretabilitas ke dalam kerangka evaluasi multi-dimensi.

1.6 Manfaat Penelitian

Hasil dari penelitian ini diharapkan dapat memberi manfaat baik secara teoritis maupun praktis:

1.6.1 Manfaat teoritis

1) Memberikan kontribusi empiris terhadap pemahaman tentang kemampuan dan keterbatasan model *pre-trained* berbasis Transformer dalam menangani bahasa daerah Indonesia yang berstatus *low-resource*, khususnya bahasa Bali.
2) Menyediakan bukti empiris mengenai efektivitas SHAP sebagai metode XAI untuk menginterpretasikan model klasifikasi sentimen pada konteks bahasa *low-resource*, yang selama ini didominasi oleh penelitian pada bahasa-bahasa *high-resource*.
3) Memperkaya literatur tentang pengaruh kualitas tokenisasi terhadap performa model pada bahasa yang tidak termasuk dalam data pelatihan utama model, melalui pendekatan audit *tokenizer* berbasis leksikon.
4) Menghasilkan kerangka evaluasi multi-dimensi (performa, tokenisasi, interpretabilitas) yang dapat direplikasi sebagai prosedur evaluasi pada penelitian sejenis, dengan penyesuaian terhadap karakteristik kosakata masing-masing dataset.

1.6.2 Manfaat Praktis

1) Memberikan rekomendasi model *pre-trained* yang paling sesuai untuk tugas analisis sentimen pada teks berbahasa Bali dalam dataset NusaX, yang dapat digunakan oleh peneliti, pengembang, maupun pemangku kepentingan di bidang pengembangan teknologi bahasa Bali.
2) Mengidentifikasi kualitas tokenizer sebagai faktor kunci yang dapat menjadi target pengembangan konkret untuk meningkatkan performa pemrosesan bahasa Bali di masa depan.
3) Mendukung upaya pelestarian bahasa Bali dalam dimensi teknologi sebagai langkah awal, yaitu dengan menyediakan bukti empiris mengenai perilaku model *pre-trained* terhadap kosakata bahasa Bali yang dapat menjadi dasar bagi pengembangan *tokenizer* maupun model bahasa khusus bahasa Bali. Kontribusi ini sejalan dengan pilar pengembangan sumber daya dan teknologi bahasa dalam Dekade Internasional Bahasa Pribumi (*International Decade of Indigenous Languages*) 2022 hingga 2032 yang dicanangkan Perserikatan Bangsa-Bangsa dengan UNESCO sebagai badan pelaksana utama (UNESCO, 2022).