BAB 1 PENDAHULUAN

1.1. Latar Belakang

Media sosial telah menjadi bagian penting dalam kehidupan sehari-hari, termasuk di Indonesia yang merupakan salah satu negara dengan jumlah pengguna X (Twitter) terbesar di dunia, dengan sekitar 27,5 juta pengguna aktif pada Oktober 2023 [1]. X, sebagai platform media sosial populer, digunakan oleh banyak orang di berbagai negara untuk berbagi pemikiran dan aktivitas melalui kalimat-kalimat singkat. Basis pengguna yang besar dan aktif ini menjadikan X tidak hanya sebagai sumber informasi yang relevan, tetapi juga sebagai peluang untuk menganalisis perilaku individu, termasuk pola kepribadian, melalui penelitian berbasis teks [16]. Meskipun data dari media sosial sangat melimpah, tantangan utama yang dihadapi adalah bagaimana mengekstrak informasi kepribadian dengan akurat dari teks yang bersifat tidak terstruktur dan tidak merata.

Kepribadian mencakup pola pikir, emosi, dan perilaku yang membedakan setiap individu [2]. Salah satu model yang sering digunakan untuk menganalisisnya adalah *Big Five Personality* (OCEAN), yang terdiri dari lima dimensi: *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, dan *Neuroticism* [3]. Model ini menawarkan cara yang komprehensif untuk memahami kepribadian dan banyak diterapkan dalam psikologi. Untuk mengukur dimensi-dimensi tersebut, biasanya digunakan kuesioner seperti BFI-44, yang diperkenalkan oleh John, Donahue, dan Kentle pada tahun 1991. BFI-44 mencakup 44 pertanyaan, di mana peserta menanggapi setiap pernyataan menggunakan skala Likert dari "sangat tidak setuju" hingga "sangat setuju", untuk menunjukkan seberapa relevan pernyataan tersebut dengan diri mereka. Kuesioner ini telah terbukti valid di berbagai kelompok budaya, menjadikannya alat yang populer dan efektif untuk menilai kepribadian [20], [21].

Namun, penerapan model *Big Five Personality* dalam analisis data media sosial menghadapi beberapa kendala. Salah satu tantangan utama dalam penelitian ini adalah keterbatasan jumlah data pengguna, yaitu hanya 381 akun dengan cuitan yang telah diberi label kepribadian. Selain jumlahnya yang kecil, distribusi label

dalam *dataset* juga tidak seimbang, di mana kelas mayoritas lebih dominan dalam prediksi, sementara kelas minoritas kurang terwakili. Ketidakseimbangan ini dapat menyebabkan bias dalam hasil klasifikasi, mengurangi akurasi model, dan membatasi kemampuan model dalam melakukan generalisasi dengan baik [22].

Untuk mengatasi tantangan ini, penelitian ini mengusulkan penggunaan dua teknik utama yaitu *Random Oversampling* (ROS) dan *Easy Data Augmentation* (EDA). ROS bertujuan untuk menyeimbangkan distribusi kelas dengan menggandakan sampel dari kelas minoritas, sehingga model dapat belajar secara lebih adil terhadap semua kelas. Sementara itu, EDA memperkaya dataset dengan teknik augmentasi teks seperti penggantian sinonim, penyisipan acak, pertukaran kata, dan penghapusan kata. Dengan menerapkan kedua teknik ini, diharapkan model dapat memiliki data yang lebih representatif dan meningkatkan akurasi klasifikasi [14], [15].

Sebagai model utama dalam penelitian ini, digunakan *A Robustly Optimized* BERT *Pretraining Approach* (RoBERTa) untuk mengklasifikasikan kepribadian *Big Five* berdasarkan data dari X. RoBERTa merupakan pengembangan dari *Bidirectional Encoder Representations from Transformers* (BERT) yang menghilangkan *Next Sentence Prediction* (NSP) dan menerapkan *dynamic masking* untuk meningkatkan variasi data yang dilihat oleh model. Dengan prosedur *pretraining* yang lebih optimal dan data yang lebih besar, RoBERTa telah terbukti unggul dalam berbagai tugas pemrosesan bahasa alami, termasuk analisis sentimen dan klasifikasi kepribadian [4].

Penelitian sebelumnya menunjukkan bahwa RoBERTa mampu melakukan tugas analisis sentimen dan identifikasi kepribadian. Beberapa contoh penerapannya termasuk penelitian oleh Muhammad Mahrus Zain et al. [5] dan Eggi Farkhan Tsani serta Derwin Suhartono [6], yang memanfaatkan RoBERTa dalam analisis sentimen dan klasifikasi kepribadian. Selain itu, penelitian lain seperti yang dilakukan oleh Yani dan Maharani [7] yang menggunakan RoBERTa untuk menganalisis konten *cyberbullying* di X, serta Putra dan Setiawan [8] yang mengaplikasikan RoBERTa dalam analisis sentimen, juga menunjukkan efektivitas metode ini dalam memproses data media sosial. Penelitian oleh Murarka et al. [9]

mengenai deteksi penyakit mental lebih lanjut membuktikan keberhasilan RoBERTa dalam menghadapi tugas klasifikasi teks yang kompleks.

Dengan menerapkan kombinasi RoBERTa, ROS, dan EDA, penelitian ini bertujuan untuk meningkatkan akurasi dalam klasifikasi kepribadian *Big Five* berbasis data media sosial X. Pendekatan ini diharapkan dapat mengatasi tantangan ketidakseimbangan data serta memberikan kontribusi dalam pengembangan metode klasifikasi teks berbasis kepribadian menggunakan kecerdasan buatan.

1.2. Rumusan Masalah

Dalam penelitian ini, terdapat beberapa permasalahan yang dirumuskan sebagai berikut:

1. Bagaimana metode RoBERTa dapat diimplementasikan untuk mengklasifikasikan kepribadian *Big Five* berdasarkan data dari media sosial X?
2. Seberapa baik performansi metode RoBERTa dalam klasifikasi kepribadian *Big Five* berdasarkan metrik evaluasi yang meliputi akurasi, *F1-score*, presisi, dan *recall*?
3. Pendekatan apa yang paling efektif untuk meningkatkan performansi model dalam klasifikasi kepribadian *Big Five* dengan menggunakan metode RoBERTa?

1.3. Tujuan dan Manfaat

Tujuan dari penelitian ini adalah sebagai berikut:

1. Mengimplementasikan metode RoBERTa dalam klasifikasi kepribadian *Big Five* berdasarkan data dari media sosial X.
2. Mengevaluasi performansi RoBERTa melalui metrik evaluasi seperti akurasi, *F1-score*, presisi, dan *recall*.
3. Mengidentifikasi pendekatan yang paling efektif dalam meningkatkan performansi model dalam klasifikasi kepribadian *Big Five*.

Sementara itu, manfaat dari penelitian ini diharapkan dapat memberikan kontribusi

sebagai referensi untuk penelitian lebih lanjut di bidang klasifikasi teks, khususnya dalam konteks klasifikasi multi-label untuk klasifikasi kepribadian.

1.4. Batasan Masalah

Penelitian ini memiliki beberapa batasan untuk menyederhanakan ruang lingkup dan memastikan penyelesaian penelitian dalam waktu yang tersedia. Batasan-batasan tersebut adalah sebagai berikut:

1. Data diperoleh melalui *crawling* cuitan dari pengguna media sosial X yang telah mengisi kuesioner *Big Five Inventory* (BFI-44). Hasil kuesioner ini digunakan sebagai label untuk lima dimensi *Big Five*, dengan total data yang terdiri dari 381 akun. Pemilihan jumlah akun ini didasarkan pada ketersediaan data yang telah mengisi kuesioner BFI-44 yang lengkap, tanpa adanya pembatasan jumlah cuitan yang diambil.
2. Sebagian besar cuitan dalam *dataset* ditulis dalam bahasa Indonesia, mengingat fokus penelitian pada pengguna media sosial X di Indonesia. Namun, beberapa cuitan mungkin menggunakan bahasa Inggris.
3. Penelitian ini mencakup lima dimensi kepribadian dari model *Big Five Personality*: *openness*, *conscientiousness*, *extraversion*, *agreeableness*, dan *neuroticism*.
4. Penelitian ini menggunakan metode RoBERTa sebagai model utama dalam klasifikasi.
5. Penelitian ini terbatas pada penggunaan *Random Oversampling* (ROS) dan *Easy Data Augmentation* (EDA) sebagai teknik untuk menangani ketidakseimbangan data dan meningkatkan performa model.
6. Penelitian ini hanya mengukur kepribadian yang nampak dari cuitan pengguna saat ini.

1.5. Metode Penelitian

Penelitian ini dilakukan melalui beberapa tahapan yang terstruktur, meliputi studi literatur, pengumpulan data, *preprocessing data*, pengembangan model, serta

eksperimen dan evaluasi. Setiap tahapan dirancang untuk meningkatkan akurasi model dan memastikan pendekatan yang efektif.

1.5.1. Studi Literatur

Penelitian ini diawali dengan mempelajari teori *Big Five Personality*, metode RoBERTa, dan teknik pengolahan data teks untuk klasifikasi multi-label. Literatur yang digunakan terdiri dari jurnal, buku, dan artikel ilmiah yang relevan, yang mendukung dasar teori dan pendekatan yang digunakan dalam penelitian ini.

1.5.2. Pengumpulan Data

Data diperoleh melalui *crawling* cuitan dari pengguna media sosial X yang telah mengisi kuesioner *Big Five Inventory* (BFI-44). Hasil kuesioner ini digunakan sebagai label untuk lima dimensi kepribadian *Big Five*: *openness*, *conscientiousness*, *extraversion*, *agreeableness*, dan *neuroticism*. *Dataset* yang dikumpulkan mencakup cuitan dalam bahasa Indonesia dan sebagian dalam bahasa Inggris, yang menggambarkan kebiasaan pengguna media sosial di Indonesia.

Penelitian ini mengambil seluruh cuitan dari 381 akun pengguna yang telah mengisi kuesioner BFI-44, tanpa membatasi jumlah cuitan yang diambil. Namun, distribusi cuitan antar akun tidak merata, yang dapat mempengaruhi akurasi model. Akun dengan banyak cuitan cenderung lebih dominan, sementara akun dengan sedikit cuitan mungkin kurang terwakili. Untuk mengatasi masalah ini, digunakan Random Oversampling (ROS) dan Data Augmentation (EDA) untuk menyeimbangkan data dan memperkaya variasinya. Meskipun demikian, ketidakseimbangan ini tetap menjadi batasan yang perlu dipertimbangkan saat menilai hasil penelitian.

1.5.3. *Preprocessing Data*

Proses *preprocessing* bertujuan untuk memastikan data memiliki kualitas tinggi, meminimalkan *noise*, dan meningkatkan efisiensi pelatihan model. Penelitian ini menggunakan dua skenario *preprocessing*:

1. *Half Preprocessing*: Meliputi *cleansing*, *case folding*, dan *tokenization*. Pada skenario ini, data hanya dibersihkan dan distandarisasi, serta diproses menjadi token.

2. *Full Preprocessing*: Meliputi *cleansing*, *case folding*, *tokenization*, *normalization*, penghapusan *stopword*, dan *stemming*. Skenario ini lebih komprehensif dan bertujuan untuk mengurangi *noise* dalam data yang dapat mengganggu proses pelatihan.

Kedua skenario preprocessing ini diuji untuk melihat pengaruh kompleksitas dan teknik preprocessing terhadap akurasi model RoBERTa. Pengujian ini bertujuan untuk mengidentifikasi proses preprocessing yang memberikan kontribusi terbesar dalam meningkatkan performansi model dalam klasifikasi kepribadian berdasarkan *Big Five*.

1.5.4. *Tuning Hyperparameter*

Setelah tahap preprocessing, dilakukan pencarian *hyperparameter* terbaik menggunakan metode *random search*. Parameter yang disesuaikan meliputi *learning rate*, *batch size*, dan *weight decay*, yang mempengaruhi performa model saat pelatihan.

Metode *random search* dipilih karena memungkinkan pencarian kombinasi parameter yang lebih efisien dan fleksibel. Proses *tuning* dilakukan terpisah pada *dataset* dengan *half preprocessing* dan *full preprocessing*, untuk menyesuaikan parameter dengan karakteristik masing-masing *dataset*. Kombinasi parameter terbaik kemudian digunakan untuk melatih model utama.

1.5.5. Pengembangan Model

Pengembangan model dilakukan menggunakan RoBERTa sebagai model utama untuk klasifikasi kepribadian *Big Five*. Beberapa pendekatan diterapkan selama pelatihan untuk meningkatkan performa model, antara lain:

1. *Baseline*: Model dilatih tanpa penyeimbangan data, sebagai acuan performa awal.
2. *Random Oversampling* (ROS): Teknik ini menyeimbangkan distribusi data antar label dengan menduplikasi data dari label minoritas hingga setara dengan mayoritas [14].

3. *Data Augmentation*: Teknik augmentasi data dilakukan untuk menambah variasi *dataset*, menggunakan metode *Easy Data Augmentation* (EDA) yang mencakup:

    - *Synonym Replacement* (SR): Mengganti kata-kata dalam teks dengan sinonimnya secara acak, untuk menambah keragaman kalimat.
    - *Random Insertion* (RI): Menyisipkan kata acak ke dalam teks untuk menambah variasi.
    - *Random Swap* (RS): Menukar posisi dua kata secara acak dalam kalimat untuk menciptakan variasi struktur tanpa merubah makna.
    - *Random Deletion* (RD): Menghapus kata-kata dalam teks dengan probabilitas tertentu, bertujuan untuk memperkenalkan variasi dalam panjang teks dan mengurangi ketergantungan model pada kata-kata tertentu.

    Penerapan teknik ini bertujuan untuk memperkaya data dan membantu model mengenali pola yang lebih beragam.

4. *Data Augmentation* + ROS: Kombinasi antara *Data Augmentation* dan *Random Oversampling* dilakukan untuk menciptakan *dataset* yang lebih beragam sekaligus seimbang.

5. Pengujian Pembagian Data

    Setiap pendekatan diterapkan pada dua skenario pembagian data:

    - 80/20: 80% data digunakan untuk pelatihan dan 20% untuk pengujian.
    - 70/30: 70% data digunakan untuk pelatihan dan 30% untuk pengujian.

Dengan kombinasi teknik yang berbeda, enam belas skenario pengujian diterapkan untuk mengidentifikasi pendekatan terbaik dalam meningkatkan performa model klasifikasi multi-label, khususnya dalam prediksi kepribadian *Big Five*.

1.5.6. Eksperimen dan Evaluasi

Eksperimen dilakukan untuk mengevaluasi kombinasi *preprocessing* dan pendekatan yang diterapkan. Evaluasi model dilakukan menggunakan metrik akurasi, *F1-score*, presisi, dan *recall*, dengan tujuan untuk mengidentifikasi pendekatan terbaik dalam meningkatkan performa model.