BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Depresi merupakan salah satu beban kesehatan global yang paling signifikan saat ini. Organisasi Kesehatan Dunia mencatat sekitar 332 juta penderita depresi di dunia, dan sekitar 727.000 kematian akibat bunuh diri pada tahun 2021 yang sebagian besarnya berkaitan dengan kondisi depresi [1]. Lebih dari 75% penderita di negara berpenghasilan menengah dan rendah, termasuk Indonesia, tidak meneriman penanganan yang memadai akibat keterbatasan tenaga profesional, stigma sosial, dan distribusi layanan yang tidak merata [1]. Kesenjangan ini bahkan lebih tinggi di Indonesia, studi berbasis data nasional menemukan bahwa hanya 9,3% penduduk dengan indikasi depresi yang memperoleh penanganan, sebagian besar terhambat oleh keterbatasan tenaga profesional dan distribusi fasilitas yang tidak merata antarwilayah [2]. Kesenjangan antara prevalensi penyakit dan ketersediaan layanan kesehatan mental ini menjadi alasan kuat untuk mengembangkan mekanisme skrining alternatif yang dapat menjangkau populasi lebih luas tanpa bergantung sepenuhnya pada kunjungan klinik konvensional [2].

Media sosial menawarkan peluang untuk skrining berbasis perilaku digital. Indonesia memiliki 143 juta pengguna media sosial aktif [3], dan secara khusus menempati peringkat keempat dunia untuk pengguna X(Sebelumnya twitter) dengan 27,1 juta pengguna pada tahun 2024 [4]. Pengguna X cenderung mengekspresikan kondisi emosional dan tekanan psikologis secara spontan dalam bentuk teks pendek, sehingga platform ini menjadi sumber data yang relevan untuk identifikasi gejala depresi.

Secara klinis, depresi pada dasarnya merupakan gangguan afektif yang dicirikan oleh suasana hati negatif yang menetap serta hilangnya minat dan kemampuan merasakan kesenangan atau *anhedonia* [5]. Kerangka tripartit yang banyak digunakan dalam psikologi klinis menjelaskan bahwa depresi ditandai oleh tingginya afek negatif seperti kesedihan, kemarahan, dan ketakutan yang berpadu dengan rendahnya afek positif seperti kegembiraan dan kasih sayang [6]. Pola afektif ini tidak hanya muncul dalam asesmen klinis, tetapi juga terefleksi dalam bahasa sehari-hari pengguna medial sosial. Frekuensi kemunculan kata beremosi negatif pada unggahan X dan Facebook terbukti berkaitan dengan tingkat keparahan

depresi [7]. Dengan demikian, emosi yang terekspresi dalam *tweet* dapat berfungsi sebagai penanda yang dapat diobservasi untuk mengindikasikan kecenderungan depresi, sehingga label emosi pada teks menjadi dasar untuk membedakan kondisi terindikasi depresi dari kondisi normal.

Untuk menangkap penanda emosional tersebut secara otomatis dari volume data yang besar, penelitian deteksi depresi memanfaatkan *Natural Language Processing*(NLP). Seiring berkembangnya pemanfaatan NLP, pendekatan deteksi depresi dari teks media sosial bergeser dari metode berbasis fitur manual menuju model *machine learning* dan *deep learning* yang mempelajari representasi bahasa secara otomatis [8]. Tinjauan sistematis dan meta-analisis terkini mengonfirmasi bahwa teks media sosial berkolerasi kuat dengan kondisi depresi dan bahwa pendekatan *deep learning* secara konsisten memberikan performa prediksi yang lebih tinggi dibandingkan metode konvensional [9, 10]. Meskipun akurasi meningkat namun transparansinya menurun, model *deep learning* bersifat *black-box*, yaitu menghasilkan prediksi tanpa mengungkapkan fitur linguistik mana yang berkontribusi terhadap klasifikasi tersebut [11].

Kurangnya mekanisme penjelasan ini menjadi masalah mendasar ketika model digunakan untuk mendukung pengambilan keputusan klinis. Bagi tenaga kesehatan, prediksi yang tidak dapat ditelusuri dasarnya tidak memenuhi standar tanggung jawab profesional dan etika praktik [12, 13]. Ketika sistem menandai seorang pengguna terindikasi depresi, klinisi membutuhkan informasi mengenai pola bahasa yang memicu klasifikasi tersebut agar dapat memvalidasi temuan dan menentukan tindak lanjut yang tepat, sehingga transparansi pada model NLP kesehatan mental menjadi kebutuhan yang konsisten ditekankan dalam penelitian terkait [14, 15].

Sebelum interpretabilitas dapat diterapkan secara bermakna, model deteksi terlebih dahulu mencapai performa klasifikasi yang memadai pada karakteristik teks X berbahasa Indonesia yang informal. Beberapa penelitian terdahulu pada deteksi depresi dari teks X (sebelumnya Twitter) berbahasa Indonesia secara konsisten menunjukkan bahwa pemilihan model bahas merupakan faktor penentu performa yang lebih dominan dibandingkan arsitektur klasifikasi semata. Pada sisi perbandingan IndoBERT dan IndoBERTweet, penggunaan IndoBERT yang dilatih pada teks formal Wikipedia dan artikel berita hanya menghasilkan akurasi 51% dengan F1-*score* 31% pada data X Indonesia [16], mengonfirmasi bahwa ketidaksesuaian antara domain *pre-training* yang formal dengan karakteristik teks X informal merupakan hambatan mendasar yang tidak dapat diatasi melalui *fine-*

tuning saja. Penerapan IndoBERTweet yaitu model *pre-trained* berbasis BERT yang dikembangkan secara khusus untuk teks X Indonesia [17] pada tugas deteksi depresi yang sama mencapai akurasi 82%, melampaui seluruh pendekatan berbasis model bahasa umum sebelumnya [18].Keunggulan IndoBERTweet pada domain ini turut diperkuat oleh studi terdahulu [19].

Selain pemilihan model bahasa, arsitektur klasifikasi sekuensial juga berperan dalam menangkap struktur kontekstual teks. BiLSTM yang memproses urutan teks secara bidireksional telah terbukti efektif untuk teks depresi berbahasa Indonesia dengan akurasi 83,46% dan F1-*score* 87,11% [20]. Keunggulan utamanya bukan pada nilai akurasi saja, namun pada kemampuan memodelkan konteks dua arah, sehingga mengatasi ketebatasan LSTM-RNN unidireksional dengan representasi kata statis yang tidak mampu menangkap makna kontekstual secara dinamis [21]. Integrasi IndoBERTweet dan BiLSTM telah diterapkan pada penelitian terdahulu dalam konteks deteksi ujaran kebencian berbahasa Indonesia dengan akurasi 93,7% dan F1-*score* 93,3%, membuktikan bahwa sinergi antara representasi kontekstual *domain-specific* IndoBERTweet dan BiLSTM menghasilkan performa yang lebih unggul dibandingkan masing masing komponen secara mandiri [22]. Kombinasi representasi BERT dengan BiLSTM juga telah diterapkan pada teks berbahasa Indonesia dan menunjukkan hasil yang menjanjikan [23, 24].

Meskipun kombinasi model bahasa *domain-specific* dan arsitektur sekuensial telah menghasilkan upaya performa klasifikasi yang tinggi, capaian tersebut belum diiringi upaya menjelaskan dasar keputusan model. Survei multibahasa terhadap dataset deteksi gangguan mental dari berbagai bahasa mengonfirmasi bahwa studi yang menerapkan *Explainable AI* (XAI) pada deteksi depresi berbahasa Indonesia masih sangat terbatas [25, 26]. Kekurangan ini membatasi kontribusi penelitian NLP kesehatan mental Indonesia terhadap praktik klinis yang dapat dipertanggungjawabkan, sekaligus mempertegas kebutuhan transparansi yang telah diuraikan sebelumnya. Di antara metode XAI *post-hoc*, LIME (*Local Interpretable Model-agnostic Explanations*) sesuai untuk kebutuhan ini karena mengaproksimasi keputusan model *black-box* menggunakan model *interpretable* sederhana secara lokal di sekitar prediksi yang ingin dijelaskan, sehingga memungkinkan identifikasi kontribusi token-token spesifik pada *tweet* individual terhadap keputusan klasifikasi [14, 27].

Oleh karena itu, dilakukan penelitian menggunakan model *hybrid* IndoBERTweet(`indolem/indobertweet-base-uncased`) yang dikombinasikan

dengan lapisan BiLSTM untuk mendeteksi depresi pada teks X berbahasa Indonesia, kemudian menerapkan LIME untuk menafsirkan keputusan model pada tingkat *tweet* individual.

1.2 Rumusan Masalah

Pada bagian ini dijabarkan permasalahan utama yang hendak diselesaikan dalam bentuk poin-poin. Penulisannya bersifat pertanyaan ilmiah yang dituliskan secara singkat namun jelas.

1. Bagaimana penerapan model *hybrid* IndoBERTweet dan BiLSTM dengan interpretabilitas LIME untuk deteksi depresi pada teks X berbahasa Indonesia?
2. Berapa tingkat akurasi, precision, recall, dan F1-Score dari model IndoBERTweet dan BiLSTM yang dibangun untuk mengklasifikasikan indikasi depresi?

1.3 Batasan Permasalahan

1. *Dataset* yang digunakan merupakan gabungan dari *Indonesian Twitter Emotion Dataset* [28], *EmoTweetID-Human* [29] dan *Emotion Dataset from Indonesian public opinion* [30], dengan analisis yang dibatasi hanya pada *tweet* berbahasa Indonesia dalam pengaturan klasifikasi biner: depresi vs. bukan depresi.
2. Pemodelan deteksi depresi secara eksklusif menggunakan arsitektur *hybrid* IndoBERTweet (`indolem/indobertweet-base-uncased`) yang dikombinasikan dengan lapisan BiLSTM; metrik evaluasi yang digunakan meliputi akurasi, presisi, *recall*, dan F1-*score*.
3. Interpretasi terhadap keputusan model dibatasi hanya menggunakan metode *Explainable AI* (XAI) berbasis LIME (*Local Interpretable Model-agnostic Explanations*), mencakup penjelasan lokal pada tingkat *tweet* individual yang divisualisasikan dalam bentuk daftar token-token kontribusi tertinggi beserta bobotnya secara *post-hoc*, arsitektur model yang secara inheren interpretatif tidak dievaluasi dalam penelitian ini.

4. Penjelasan LIME merupakan aproksimasi lokal yang dihasilkan melalui perturbasi acak pada teks masukan, sehingga sensitif terhadap jumlah sampel perturbasi dan dapat menghasilkan penjelasan yang sedikit berbeda pada eksekusi berulang. Selain itu, LIME tidak menetapkan hubungan kausal antara token linguistik dan keberadaan depresi, temuan interpretabilitas merupakan observasi analitis dan tidak dapat dijadikan kriteria diagnostik klinis maupun penilaian psikologis.
5. Penelitian ini tidak mencakup penerapan sistem secara *real-time* maupun implementasi pada level produksi.

1.4 Tujuan Penelitian

1. Menerapkan model *hybrid* IndoBERTweet dan BiLSTM dengan interpretabilitas LIME untuk deteksi depresi pada teks X berbahasa Indonesia.
2. Mengukur tingkat akurasi, *precision, recall*, dan *F1-score* dari model IndoBERTweet dan BiLSTM yang dibangun untuk mengklasifikasikan indikasi depresi.

1.5 Manfaat Penelitian

1. Memberikan kontribusi literatur dan referensi bagi penelitian selanjutnya di bidang *Natural Language Processing*(NLP), khususnya mengenai efektivitas arsitektur *hybrid* IndoBERTweet dan BiLSTM untuk klasifikasi teks informal di media sosial
2. Memperkaya kajian *Explainable AI* (XAI) pada bahasa yang masih kurang terwakili (*underrepresented language*) dengan menerapkan LIME untuk menjelaskan model deteksi depresi berbahasa Indonesia.
3. Menjadi landasan awal dalam pengembangan *prototype* sistem *screening* kesehatan mental digital yang tidak hanya akurat, tetapi juga transparan dan dapat diinterpretasikan, sehingga visualisasi dari LIME dapat membantu tenaga profesional dalam memahami pola bahasa pemicu klasifikasi depresi.

1.6 Sistematika Penulisan

Sistematika penulisan laporan penelitian ini disusun dalam lima bab sebagai berikut.

1. **Bab 1 PENDAHULUAN**

   Menguraikan konteks dan latar belakang penelitian mengenai deteksi depresi dari teks X berbahasa Indonesia menggunakan model IndoBERTweet dan BiLSTM yang dilengkapi analisis interpretabilitas berbasis LIME. Bab ini juga mencakup rumusan masalah, batasan penelitian, tujuan, dan manfaat penelitian.

2. **Bab 2 LANDASAN TEORI**

   Menyajikan landasan teori yang mendasari penelitian, meliputi konsep klinis depresi dan relevansinya dengan media sosial, dasar-dasar NLP dan praproses teks, arsitektur *Transformer* dan BERT, model IndoBERTweet, komponen BiLSTM, integrasi keduanya dalam arsitektur *hybrid*, kerangka XAI dan metode LIME.

3. **Bab 3 METODOLOGI PENELITIAN**

   Menjelaskan metodologi penelitian secara sistematis, mulai dari pengumpulan dan praproses data, perancangan dan pelatihan model IndoBERTweet + BiLSTM, hingga prosedur analisis interpretabilitas menggunakan LIME.

4. **Bab 4 HASIL DAN DISKUSI**

   Menyajikan hasil eksperimen berupa metrik evaluasi model klasifikasi dan temuan analisis LIME pada tingkat lokal, disertai diskusi komparatif terhadap penelitian terdahulu yang relevan.

5. **Bab 5 KESIMPULAN DAN SARAN**

   Merangkum simpulan dari seluruh penelitian dan memberikan saran untuk penelitian lanjutan di masa depan.