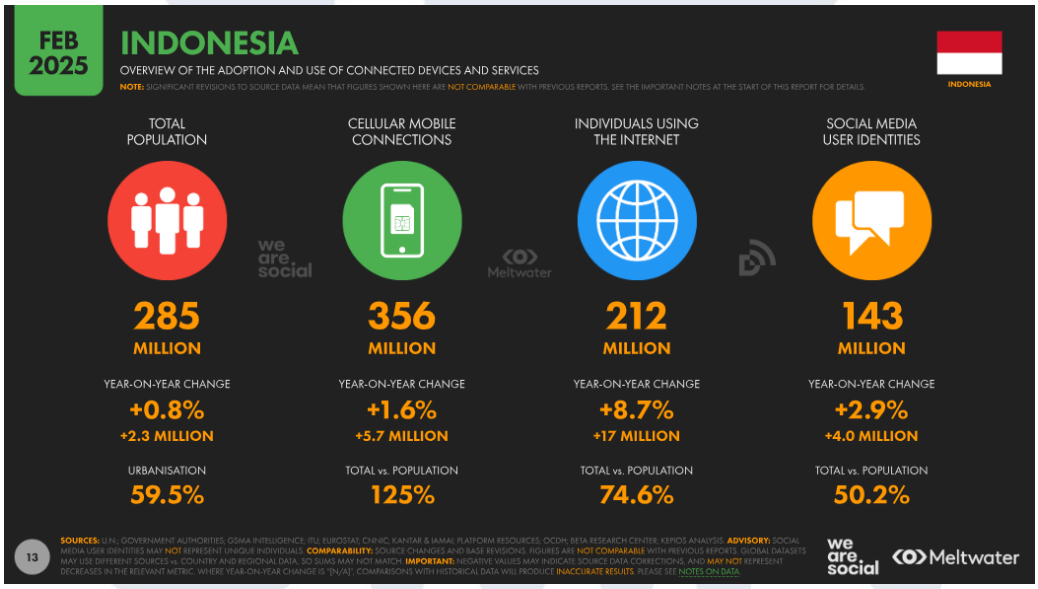
BAB I

PENDAHULUAN

1.1 Latar Belakang

Media sosial semakin berkembang dan maju secara pesat dalam kehidupan manusia [1]. Platform media sosal telah terintegrasi dengan berabgai aktivitas manusia sehari-hari seperti pendidikan, bisnis, hiburan sehingga menjadikan media sosial bagian yang tidak terpisahkan dari kebutuhan hidup sehari-hari manusia [2]. Hal tersebut dapat dilihat pada gambar 1.1 yang menunjukkan jumlah pengguna media sosial di Indonesia mencapai setengah dari populasi [3]. Namun seiring dengan meningkatnya adopsi teknologi dan tingginya intensitas penggunaan media sosial, muncul permasalahan atau ancaman baru berupa *cyberbullying* yang berpotensi membahayakan keselamatan dan kenyamanan pengguna [2].

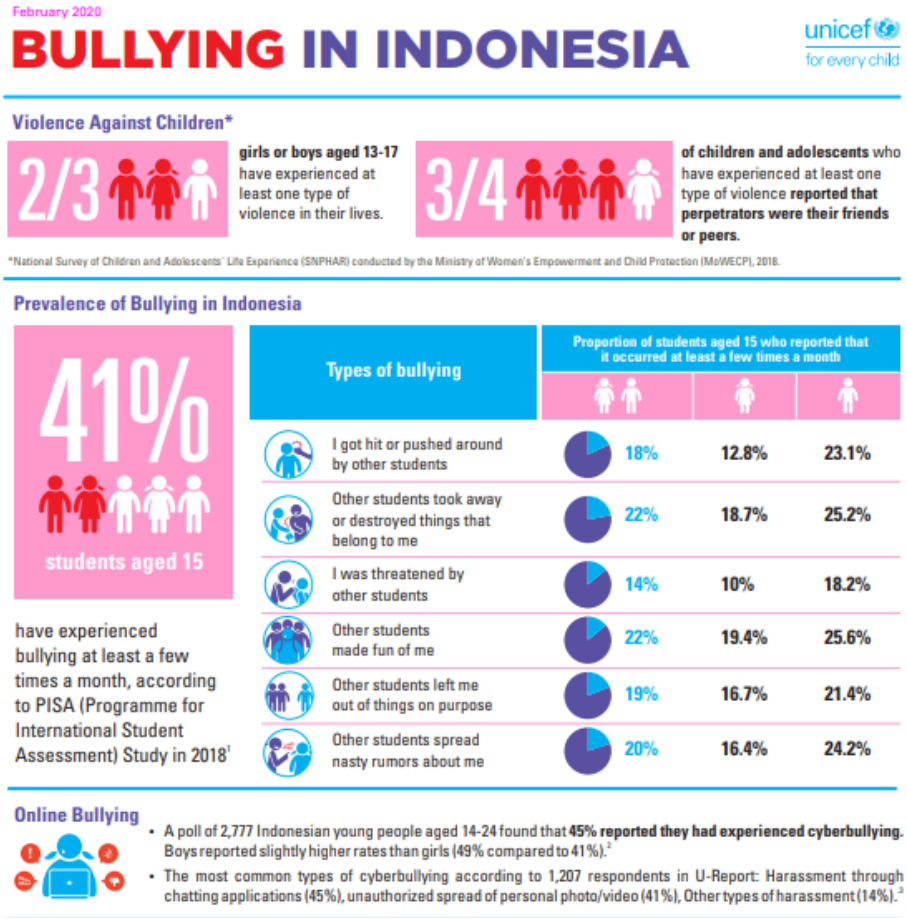


Gambar 1. 1 Statistik Digital Indonesia [3]

Gambar 1.1 menunjukkan bahwa negara Indonesia memiliki total populasi sebesar 285 juta manusia dengan tingkat urbanisasi sebesar 59,5% [3]. Jumlah pengguna media sosial di Indonesia juga sudah mencapai sebesar 143 juta manusia atau sama dengan 50,2% dari populasi manusia di Indonesia [3]. Penggunaan media sosial di Indonesia terjadi banyak masalah, salah satu di antaranya adalah *cyberbullying [4]*. *Cyberbullying* biasanya dilakukan melalui platform digital

seperti seperti forum *online*, media sosial, permainan secara *online* atau daring yang memungkinkan individu dari berbagai belahan dunia untuk berinteraksi, berkumpul dan menyebarkan informasi palsu dalam satu ruang maya atau digital [2]. Fenomena *cyberbullying* telah menjadi isu kesehatan mental yang serius, mengingat tingginya risiko korban untuk melakukan tindakan bunuh diri [5]. Hal ini disebabkan oleh dampak jangka panjang dari *cyberbullying* terhadap perkembangan psikologis individu terutama dalam masa pertumbuhan hingga dewasa [6]. *Cyberbullying* juga dapat memberikan ketidaknyamanan secara emosional baik dalam jangka pendek maupun panjang setelah menjadi sasaran serangan siber ini [7].

Pada tahun 2020, World Health Organization (WHO) melakukan penelitian kasus *cyberbullying* dari 45 negara terhadap 227.441 remaja berusia 15, 13 dan 11. Sebanyak 6% remaja mengakui telah melakukan *cyberbullying* atau perundungan dan 10% remaja mengakui pernah mengalami *cyberbullying* sebanyak tiga atau dua kali dalam rentang waktu dua bulan [8]. WHO juga melakukan penelitian pada negara Indonesia yang dapat dilihat pada gambar 1.2 menunjukkan sebesar 45% remaja Indonesia dengan rentang umur 14 hingga 24 tahun mengalami kasus *cyberbullying,* korban anak laki-laki sebanyak 45% dan perempuan sebanyak 41%. Hal tersebut menandakan korban anak laki-laki lebih banyak [4]



Gambar 1. 2 Statistik Bullying di Indonesia [4]

Gambar 1.2 menunjukkan bahwa 41% siswa berusia 15 tahun yang berada di Indonesia mengaku pernah mengalami *bullying* setidaknya beberapa kali dalam sebulan baik dalam bentuk kekerasan fisik verbal maupun psikologis [4]. Bentuk *bullying* yang paling umum dialami sisa mencakup ejekan dan penghinaan sebesar 25,6% [4]. Survei terhadap 2.777 remaja yang berada di Indonesia dengan rentang umur sekitar 14 hingga 24 tahun mengungkapkan bahwa 45% dari remaja Indonesia pernah mengalami *cyberbullying* dengan persentase lebih tinggi pada laki-laki sebesar 49% dibanding perempuan sebesar 41% [4]. Bentuk *cyberbullying* yang paling umum yang terjadi pada remaja di Indonesia meliputi pelecehan melalui aplikasi *chat* sebesar 45%, penyebaran foto atau video pribadi tanpa izin sebesar 41% dan bentuk pelecahan lainya sebesar 41% [4].

Sekitar 1,5 juta manusia di seluruh dunia mengalami kasus *cyberbullying* setiap harinya dan lebih dari 87% remaja pengguna internet merasa pernah menjadi

korban perilaku tersebut. Laporan dari lembaga amal nasional anti-*bullying* menyebutkan bahwa dua dari tiga individu berusia antara 13 hingga 22 tahun pernah mengalami *cyberbullying* dalam berbagai bentuk [7]. Sebanyak 93% korban mengalami perundungan dalam berbagai bentuk baik secara tatap muka maupun melalui platform digital. Sebagian besar dari mereka tidak memiliki kemampuan untuk melawan atau menghindari kondisi yang menekan dan mengintimidasi tersebut [8]. Negara India mengalami penigkatan *cyberbullying* sebesar 6,4% pada laki-laki dan 3,8% pada perempuan. Sebanyak 16% laki-laki dan 33% perempuan mengalami gangguan depresi serta terdapat juga kasus percobaan bunuh diri yang mencapai 2,3% pada laki-laki dan 7,3% pada perempuan [6]. Dampak dari *cyberbullying* tergolong sangat serius dengan risiko tinggi terhadap tindakan bunuh diri di kalangan korban. Beberapa studi juga mengindikasikan bahwa korban dapat mengalami berbagai gangguan kesehatan mental termasuk kecemasan, depresi, hilangnya rasa percaya diri, keinginan untuk mengakhiri hidup, penurunan kinerja hingga munculnya gangguan psikis lainnya[9].

Dalam upaya menangani permasalahan ini, berbagai metode telah dikembangkan untuk mendeteksi teks *cyberbullying* secara otomatis. Algoritma SVM mendapatkan hasil akurasi sebesar 74,50% dengan presisi 74%, recall 74% dan F1 Score 74% [10]. Algoritma *Naive Bayes* mendapatkan akurasi sebesar 76%, untuk mengkalsifikasi kelas *cyberbullying* dan yang bukan *cyberbullying*[11]. Algoritma *Logistric Regression* mendapatkan akurasi rata-rata sebesar 75.9% [12]. Algoritma BiLSTM mendapatkan akurasi sebesar 80.25% untuk mengklasifikasikan kelas multi-label [13]. Algoritma BERT-BiLSTM mendapatkan akurasi sebesar 83.10% [14]. Algoritma CNN dengan BiGRU mendapatkan akurasi sebesar 78.50% sedangkan Bi-LSTM sebesar 78.74% [15]. Meskipun demikian, performa dari berbagai pendekatan tersebut masih memiliki keterbatasan dalam hal kualitas prediksi terutama pada konteks data berbahasa Indonesia. Hal ini menunjukkan bahwa masih terdapat ruang untuk optimalisasi lebih lanjut dalam pengembangan model deteksi *cyberbullying* yang lebih akurat dan adaptif terhadap karakteristik data sosial media.

Berdasarkan permasalahan tersebut, dibutuhkan sebuah sistem otomatis yang mampu mendeteksi dan mengklasifikasikan berbagai bentuk ujaran kebencian secara bersamaan dalam satu teks khususnya dalam Bahasa Indonesia yang kaya akan konteks informal dan ekspresi lokal. Peneliti mengambil topik *cyberbullying* karena keprihatinan terhadap maraknya kekerasan verbal dan ujaran kebencian di media sosial yang dapat berdampak serius pada kesehatan mental korban. Dalam penelitian ini, digunakan pendekatan hybrid antara IndoBERT dan BiLSTM. IndoBERT dipilih karena kemampuannya memahami konteks linguistik Bahasa Indonesia secara mendalam, sedangkan BiLSTM digunakan untuk menangkap urutan kata secara dua arah. Penelitian ini tidak hanya merancang sistem klasifikasi otomatis terhadap berbagai bentuk *cyberbullying* dalam Bahasa Indonesia secara multi-label, tetapi juga menganalisis performa model *deep learning* kombinasi IndoBERT dan BiLSTM dalam mengklasifikasikan 12 kategori *cyberbullying*. Evaluasi performa ini penting untuk mengetahui sejauh mana model dapat menangkap pola kekerasan verbal dalam konteks media sosi al yang dinamis dan penuh *noise*.

1.2 Rumusan Masalah

Berdasarkan pada uraian latar belakang di atas, penyelesaian yang ingin dicapai dari permasalahan tersebut melalui penelitian ini adalah sebagai berikut:

1. Bagaimana cara membangun model deteksi otomatis yang mampu mengidentifikasi berbagai bentuk *cyberbullying* dalam teks berbahasa Indonesia?
2. Bagaimana performa model *deep learning* berbasis IndoBERT dan Bi-LSTM dalam mengklasifikasikan 12 kategori *cyberbullying*?

1.3 Batasan Masalah

Terdapat beberapa batasan masalah terhadap penelitian deteksi teks *cyberbullying* secara otomatis pada sosial media, di antara lain :

1. Berfokus pada dataset teks berbahasa Indonesia
2. Pe hanya menggunakan dataset berasal dari Twitter atau X saja.

1.4 Tujuan dan Manfaat Penelitian

1.4.1 Tujuan Penelitian

Berdasarkan rumusan masalah tersebu, tujuan yang ingin dicapai dari penelitian ini ialah:

1. Membangun model deteksi otomatis *cyberbullying* berbahasa Indonesia dengan kemampuan klasifikasi multi-label untuk mengenali berbagai bentuk kekerasan verbal seperti hinaan personal, rasisme, seksisme hingga ujaran kebencian berbasis agama.
2. Mengetahui performa model *deep learning* berbasis IndoBERT dan Bi-LSTM dalam mengklasifikasikan 12 kategori *cyberbullying*.

1.4.2 Manfaat Penelitian

Manfaat yang ingin diberikan melalui penelitian ini adalah sebagai berikut:

1. Memberikan kontribusi terhadap literatur pemrosesan bahasa alami (NLP) dengan pendekatan hybrid IndoBERT dan BiLSTM untuk klasifikasi ujaran kebencian.
2. Memberikan gambaran awal melalui prototipe sistem deteksi *cyberbullying* berbasis web yang dapat digunakan untuk menunjukkan potensi penggunaan teknologi dalam mendeteksi kekerasan verbal secara otomatis.

1.5 Sistematika Penulisan

Sistematika penulisan pada penelitian ini diuraikan sebagai berikut:

BAB 1. PENDAHULUAN

Bab 1 menjelaskan mengenai latar belakang masalah yang menjadi dasar pemilihan topik penelitian, mencakup fenomena *cyberbullying* di media sosial dan dampaknya. Selanjutnya, bab ini merumuskan masalah-masalah yang akan dijawab dalam penelitian serta menjabarkan tujuan dan

manfaat yang diharapkan dari penelitian ini baik dari sisi kontribusi ilmiah maupun implementasi praktis.

BAB 2. LANDASAN TEORI

Bab 2 membahas kajian teori dan penelitian terdahulu yang relevan dengan topik penelitian. Pembahasan diawali dengan studi pustaka yang mengulas berbagai penelitian sebelumnya terkait deteksi ujaran kebencian dan analisis sentimen menggunakan pendekatan *deep learning* serta menyoroti keterbatasan yang coba diatasi dalam penelitian ini. Selanjutnya dijelaskan teori-teori dasar seperti konsep analisis sentimen, karakteristik platform Twitter/X serta definisi dan klasifikasi *cyberbullying*. Bab ini juga menguraikan kerangka kerja CRISP-DM sebagai metodologi yang digunakan, disertai pembahasan mendalam mengenai algoritma pemrosesan bahasa alami (NLP), arsitektur Bidirectional Long Short-Term Memory (BiLSTM), model pralatih IndoBERT serta metrik evaluasi seperti confusion matrix. Terakhir dijelaskan juga alat dan perangkat lunak yang digunakan dalam penelitian yaitu Google Colaboratory sebagai platform komputasi berbasis cloud dan Python sebagai bahasa pemrograman utama.

BAB 3. METODOLOGI PENELITIAN

Bab 3 Bab 3 menjelaskan tahapan-tahapan penelitian yang dilakukan berdasarkan metodologi CRISP-DM secara menyeluruh. Pembahasan dimulai dari gambaran umum objek penelitian yang berfokus pada deteksi *cyberbullying* di platform Twitter menggunakan 13.218 tweet berbahasa Indonesia dengan pendekatan multi-label classification dan algoritma deep learning BiLSTM. Selanjutnya dijelaskan metode penelitian yang mencakup enam tahap CRISP-DM dimulai dari *business understanding* terkait pemahaman masalah *cyberbullying* dan tujuan membangun sistem deteksi otomatis, kemudian tahap *data understanding* yang meliputi pengumpulan data melalui web scraping serta eksplorasi data untuk memahami distribusi dan korelasi antar label. Pada tahap *data preparation* dilakukan serangkaian proses pembersihan teks seperti *case*

folding, penghapusan emoji, mention, URL, simbol, karakter berulang, serta normalisasi kata, stopword removal, *stemming* dengan Sastrawi hingga filtering rare words. Tahap modeling mencakup pengembangan sistem klasifikasi multi-label menggunakan arsitektur hybrid IndoBERT dan BiLSTM, tokenisasi, pembagian dataset serta penerapan Focal Loss dan threshold berbasis ROC. Evaluasi dilakukan menggunakan metrik seperti *classification report* (*precision, recall, F1-score*), *confusion matrix*, dan ROC-AUC. Hasil model kemudian di-deploy ke dalam aplikasi web berbasis Gradio dan dihosting melalui Hugging Face Spaces agar dapat diakses publik. Bab ini juga mencakup teknik pengumpulan data menggunakan Tweet Harvest dan kata kunci tertentu, serta teknik analisis data berbasis pendekatan data mining untuk mengklasifikasikan teks secara sistematis dan terstruktur..

BAB 4. ANALISIS DAN HASIL PENELITIAN

Bab 4 menjelaskan hasil eksperimen dan evaluasi model secara sistematis berdasarkan tahapan CRISP-DM. Pembahasan dimulai dari analisis ulang permasalahan dan tujuan penelitian *(business understanding)*, diikuti visualisasi hasil scraping dan EDA seperti distribusi label, korelasi antar label, panjang teks, dan WordCloud *(data understanding)*. Tahap *data preparation* mencakup pembersihan data, stemming, identifikasi rare words, sinkronisasi label, serta finalisasi dataset. EDA pasca-cleaning juga ditampilkan untuk melihat perubahan distribusi data. Modeling menjelaskan pelatihan dengan dataset auto-labeled menggunakan arsitektur IndoBERT dan BiLSTM serta penerapan Focal Loss. Evaluasi dilakukan menggunakan threshold ROC-AUC, *classification report*, *confusion matrix*, dan ROC-AUC per label. Bab ini juga memuat hasil pengujian pengguna terhadap prototipe sistem yang telah dideploy sebagai aplikasi web deteksi *cyberbullying*..

BAB 5. SIMPULAN DAN SARAN

Bab 5 berisi kesimpulan dari penelitian yang telah dilakukan, termasuk pencapaian model dan implementasi aplikasi. Bab ini juga memuat saran-saran untuk pengembangan lebih lanjut, baik dari sisi teknis model maupun penerapan sistem dalam skala yang lebih luas.

.