BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Penelitian ini secara khusus memfokuskan objek kajiannya pada fenomena *cyberbullying* yang terjadi di platform media sosial Twitter. Twitter dipilih sebagai fokus penelitian ini karena karakteristiknya sebagai media sosial yang terbuka dan memiliki intensitas interaksi sangat tinggi, sehingga dikenal rawan terjadi perilaku *cyberbullying* terhadap remaja *[72]*. Kebebasan berekspresi dan kecepatan arus informasi di Twitter memungkinkan pengguna menyebarkan ujaran kebencian, hinaan, atau bentuk kekerasan verbal lain secara luas dalam waktu singkat [73]. Hal ini diperkuat oleh temuan bahwa Twitter berkontribusi paling tinggi dalam penyebaran konten bermuatan kebencian dibanding platform media sosial lain. Dengan jutaan pengguna aktif dan sifat interaksi yang spontan, Twitter menyediakan ruang ekspresi yang bebas namun rentan disalahgunakan untuk perundungan daring [35], [36], [37]. Kondisi inilah yang menjadikan Twitter relevan sebagai objek kajian utama dalam pengembangan sistem deteksi otomatis cyberbullying berbasis teks.

Penelitian ini bertujuan untuk mendeteksi dan mengklasifikasikan berbagai bentuk *cyberbullying* secara otomatis berdasarkan analisis teks cuitan (tweet) berbahasa Indonesia. Data yang digunakan dalam penelitian ini merupakan kumpulan tweet yang telah dikategorikan sebelumnya oleh peneliti terdahulu ke dalam dua belas jenis *cyberbullying* yaitu kekerasan verbal (*abusive language*), ujaran kebencian (*hate speech*), penghinaan terhadap individu maupun kelompok, diskriminasi berbasis agama, ras, dan gender, serta bentuk kekerasan lainnya seperti *body shaming*, pelecehan fisik hingga ujaran kebencian berdasarkan tingkat intensitas, yaitu lemah (*weak*), sedang (*moderate*) dan kuat (*strong*). Total data yang digunakan dalam penelitian ini berjumlah 13.218 tweet, seluruhnya berbahasa Indonesia dan telah melalui proses pelabelan manual yang divalidasi.

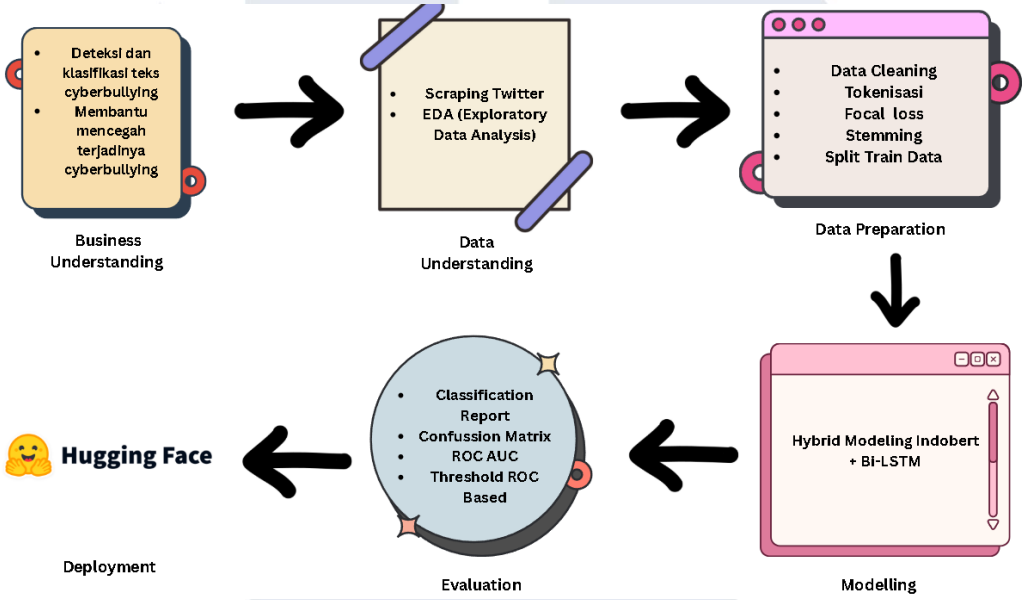
Proses klasifikasi dilakukan secara multi-label dengan pendekatan *deep learning* menggunakan arsitektur Bidirectional Long Short-Term Memory (Bi-

LSTM). Dengan menjadikan Twitter sebagai sumber data utama dan pendekatan Bi-LSTM sebagai metode pemodelan, penelitian ini berupaya membangun sistem pendeteksi *cyberbullying* otomatis yang mampu mengenali tidak hanya keberadaan kekerasan digital dalam teks, tetapi juga mengidentifikasi jenis-jenisnya secara lebih terstruktur dan komprehensif.

3.2 Metode Penelitian

Metode penelitian yang digunakan dalam studi ini adalah CRISP-DM (Cross Industry Standard Process for Data Mining), sebuah kerangka kerja standar yang banyak digunakan dalam pengembangan sistem berbasis data mining dan machine learning. Metodologi ini terdiri dari enam tahapan utama yang terintegrasi: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment. Pendekatan ini bersifat iteratif dan fleksibel, sehingga memungkinkan proses pengembangan berjalan secara sistematis meskipun dilakukan dalam skala terbatas oleh individu atau tim kecil. Pemilihan CRISP-DM dalam penelitian ini disesuaikan dengan kebutuhan dan karakteristik pengguna awal dari sistem yang dibangun, yaitu remaja yang menjadi target uji coba prototipe deteksi *cyberbullying* berbasis web. Mereka merupakan kelompok yang rentan terhadap paparan cyberbullying di media sosial dan memiliki kepentingan langsung terhadap isu ini, sehingga sangat relevan untuk dilibatkan sebagai pengguna awal sistem. Melalui keterlibatan mereka, sistem dapat diuji dari sisi fungsionalitas dan kemudahan penggunaan, serta memberikan umpan balik awal terhadap efektivitas model dalam mendeteksi kekerasan verbal secara otomatis. Seluruh tahapan dalam CRISP-DM telah selaras dengan proses penelitian ini mulai dari eksplorasi dan pemahaman data Twitter, pembersihan teks informal, tokenisasi, hingga tahap modeling multi-label dengan pendekatan hybrid IndoBERT dan BiLSTM. Model yang dihasilkan kemudian dievaluasi menggunakan metrik F1-score dan ROC-AUC, sebelum akhirnya di-deploy dalam bentuk prototipe yang dapat diakses oleh pengguna melalui platform Hugging Face. Dengan menggunakan CRISP-DM, proses pengembangan menjadi lebih terarah, terstruktur, dan memungkinkan hasil akhir dapat digunakan langsung oleh pengguna nyata dalam hal ini rekan-rekan

pelajar sebagai langkah awal dalam mengedukasi dan meningkatkan kesadaran terhadap bahaya *cyberbullying*. Berikut ini merupakan alur CRISP-DM dan tabel perbandingan pemilihan *framework* CRISP-DM, SEMMA dan KDD[74], [75] yang digunakan pada penelitian ini :



Gambar 3.1 Diagram CRISP-DM

Tabel 3. 1 Perbandingan Framework Metodologi

Indikator	CRISP-DM	SEMMA	KDD
Tahapan	*1. Business Understanding* *2. Data Understanding* *3. Data Preparation* *4. Modeling* *5. Evaluation* *6. Deployment*	*1. Sample* *2. Explore* *3. Modify* *4. Model* *5. Assessment*	*1. Pre-KDD* *2. Selection* *3. Transformation* *4. Data Mining* *5. Interpretation* *6. Post-KDD*
Kelebihan	Komprehensif, mencakup seluruh siklus proyek data mining dari tahap pemahaman bisnis	Fokus pada data dan modeling, merupakan proses iteratif yang menekankan langkah-langkah teknis	Kerangka konseptual luas, menawarkan proses penemuan pengetahuan yang menyeluruh dari

Indikator	CRISP-DM	SEMMA	KDD
	hingga deployment, sehingga dianggap lebih lengkap dibanding kerangka lain. Selain itu, prosesnya iteratif dan tahapannya dapat diulang atau dibalik (reversible) untuk memperbaiki kesalahan tanpa harus mengulang seluruh siklus dari awal.	(sampling data, eksplorasi, transformasi, modeling, dan penilaian model) sehingga sederhana dan langsung pada inti analisis. Kerangka ini efektif untuk eksperimen data mining karena mendukung berbagai teknik dan memandu pengolahan data hingga model siap dievaluasi.	seleksi data hingga ekstraksi pola. KDD bersifat iteratif dan bahkan interaktif, artinya proses dapat diulang-ulang dengan umpan balik secara terus-menerus hingga mendapatkan insight yang diinginkan. Fleksibilitas ini memungkinkan penyesuaian pada berbagai domain dan kebutuhan analisis.
Kekurangan	Metodologi ini belum mencakup aspek pendukung seperti kontrol, monitoring, komunikasi, manajemen pengetahuan, serta pemeliharaan model pasca-deployment. Selain itu, CRISP-DM dinilai terlalu panjang dan tidak fleksibel, karena tidak menyediakan panduan manajemen perubahan tim maupun model evaluasi proses yang berkelanjutan.	SEMMA memiliki cakupan terbatas karena tidak memberikan panduan detail per fase dan mengabaikan tahap pemahaman bisnis. Metodologi ini juga kurang menekankan deployment karena awalnya dirancang khusus untuk SAS Enterprise Miner, sehingga tidak ideal untuk proyek data mining end-to-end.	KDD kurang terstruktur karena tidak menjabarkan tugas spesifik di tiap tahap dan tidak memberikan panduan jelas untuk evaluasi hasil, interpretasi model, atau visualisasi. Pendekatannya yang abstrak membuat implementasinya kurang praktis dibanding kerangka seperti CRISP-DM.

3.2.1 Business Understanding

Tahapan *Business Understanding* merupakan langkah awal dalam metodologi CRISP-DM yang bertujuan untuk memahami permasalahan yang sedang terjadi dan menjadikan hal tersebut menjadi pondasi penelitian ini untuk dilakukan. Isu yang menjadi fokus utama adalah meningkatnya kasus *cyberbullying*

di media sosial. Platform ini sering menjadi tempat munculnya berbagai bentuk ujaran kebencian yang ditujukan kepada individu, kelompok atau identitas tertentu yang berdampak pada kondisi psikologis korban.

Penelitian ini bertujuan untuk membangun model berbasis *deep-learning* yang mampu melakukan deteksi secara otomatis terhadap teks yang mengandung unsur *cyberbullying* serta mengklasifikasikannya ke dalam berbagai kategori secara bersamaan atau biasa disebut klasifikasi multilabel. Hasil akhir dari penelitian ini diharapkan dapat digunakan oleh peneliti lain sebagai alat bandu dalam memetakan dan mengintervensi sebaran *cyeberbullying* secara lebih efisien.

3.2.2 Data Understanding

Tahapan *data understanding* dilakukan untuk memperoleh pemahaman yang mendalam terhadap struktur, karakteristik dan potensi tantangan yang terdapat dalam dataset yang digunakan. Pada penelitian ini, data yang digunakan berasal dari dua sumber, yang pertama data yang sudah terlabeli dari peneliti lain dan data yang belum terlabeli berasal dari kumpulan tweet yang telah di ambil oleh peneliti menggunakan alat tweet harvest. Dataset label memiliki beberapa kategori atau jenis ujaran kebencian yang mencakup kekerasan verbal terhadap individu maupun kelompok berdasarkan ras, agama, gender dan bentuk lainya seperti penghinaan fisik, merendahkan dan berkata kasar atau *toxic.* Dataset tersebut memiliki kolom multilabel ditandai dengan “HS”, “Abusive” “HS_Individual”, “HS_Group” , “HS_Religion”, “HS_Gender”, “HS_Race”, “HS_Physical”, “HS_Gender”, “HS_Other”, “HS_Weak”, “HS_Moderate” dan “HS_Strong”. Peneliti mengambil sebanyak 19.000 data dari hasil penggabungan 3 kali scraping data twitter dengan pembagian setiap satu kali pengambilan data dari twitter, target data yang dikumpuilkan sebanyak 5000 data, akan tetapi pada saat penarikan terakhir kali peneliti hanya mendapatkan 4000 data saja dikarenakan koneksi yang buruk pada saat penarikan terakhir kali.

Proses ekplorasi data juga dilakukan dengan menghitung distribusi label untuk mengetahui proporsi masing-masing kategori dalam dataset. Visualisasi seperti *bar chart* dan *heatmap* digunakan untuk mengindentifikasi porensi ketidakseimbangan data serta menggambarkan relasi atau korelasi antara jenis-jenis

cyberbullying. Selain itu dilakukan juga analisis statistik sederhana terhadap panjang teks serta pembuatan *wordcloud* untuk memahami dominasi kosakata dalam tweet yang terdapat dalam dataset. Tahapan *data understanding* penting dilakukan untuk membantu peneliti memahami data yang digunakan untuk mendeteksi dan mengklasifikasi teks *cyberbullying.* Adapun penjelasan terkait kelas atau label yang ada pada dataset diantara lain :

Tabel 3.2 Keterangan Masing-Masing Label[73]

Label	Keterangan
HS	Label ujaran kebencian
Abusive	Label bahasa kasar
HS_Individual	Label ujaran kebencian yang ditujukan terhadap individu
HS_Group	Label ujaran kebencian yang ditujukan terhadap kelompok
HS_Religion	Label ujaran kebencian yang berkaitan dengan agama atau keyakinan
HS_Race	Label ujaran kebencian yang berkaitan dengan ras atau etnis
HS_Physical	Label ujaran kebencian yang berkaitan dengan fisik atau keterbatasan
HS_Gender	Label ujaran kebencian yang berkaitan dengan gender atau jenis kelamin
HS_Other	Label ujaran kebencian yang berkaitan dengan hinaan atau fitnah lain
HS_Weak	Label ujaran kebencian ringan
HS_Moderate	Label ujaran kebencian sedang
HS_Strong	Label ujaran kebencian berat

3.2.3 Data Preparation

Proses data preparation merupakan tahapan krusial dalam penelitian ini untuk memastikan bahwa data mentah dari media sosial Twitter dapat dikonversi menjadi format yang bersih, terstruktur, dan layak digunakan dalam pelatihan model klasifikasi multi-label berbasis deep learning. Mengingat karakteristik data Twitter yang bersifat informal, padat noise, serta penuh variasi gaya bahasa dan

ekspresi, diperlukan serangkaian proses pembersihan dan transformasi teks yang sistematis. Seluruh proses ini dilakukan secara bertahap menggunakan Jupyter Notebook sebagai lingkungan pemrograman interaktif untuk eksperimen dan pengujian preprocessing pipeline.

Langkah pertama dimulai dengan case folding, yaitu mengubah seluruh huruf menjadi huruf kecil serta menormalkan karakter khusus seperti "é" menjadi "e", agar setiap kata diidentifikasi secara seragam dan tidak muncul sebagai token berbeda. Kemudian dilakukan penghapusan noise, seperti tag mention (@user), tautan (URL), simbol, angka, serta karakter non-alfabetik yang tidak memberikan kontribusi semantik terhadap konteks isi tweet. Proses ini dibantu oleh fungsi clean_text yang dirancang khusus untuk menyaring unsur-unsur tidak relevan.

Selanjutnya diterapkan reduksi karakter berulang menggunakan fungsi reduce_repeated_chars, untuk menangani ekspresi berlebihan seperti “hahahaha” atau “tolooooooong” menjadi bentuk standar seperti “haha” atau “tolong”. Langkah ini penting agar gaya ekspresi informal tidak mengganggu konsistensi tokenisasi dan tetap mempertahankan makna emosional yang ingin disampaikan oleh penulis tweet.

Proses normalisasi kata dilakukan dengan memanfaatkan lima jenis kamus dalam format JSON, yaitu kamus slang, alay, singkatan, normalisasi, dan gabungan slang. Tahap ini bertujuan untuk mengkonversi berbagai variasi kata tidak baku ke bentuk baku Bahasa Indonesia yang lebih mudah dipahami oleh model. Contohnya, kata-kata seperti "gak", "ga", dan "nggak" diseragamkan menjadi “tidak”, sedangkan "ak", "gue", atau "gw" menjadi “saya”. Dengan menyatukan ragam penulisan informal ke bentuk standar, model menjadi lebih mudah mengenali dan memahami makna dari teks yang dimasukkan.

Setelah normalisasi, dilakukan penghapusan stopword, yaitu kata-kata umum yang tidak memiliki nilai informatif tinggi dalam proses klasifikasi seperti “yang”, “dan”, “di”, dan sejenisnya. Dengan menghapus stopword, model dapat lebih fokus pada kata-kata yang memiliki bobot semantik yang tinggi dalam menentukan keberadaan cyberbullying.

Tahap berikutnya adalah stemming, yang dilakukan menggunakan library Sastrawi. Proses ini mengembalikan kata berimbuhan ke bentuk dasarnya, misalnya "memukul", "pukulan", atau "dipukul" menjadi "pukul". Tujuan stemming adalah untuk mengurangi kompleksitas morfologis Bahasa Indonesia dan menyederhanakan fitur agar model dapat mengenali entitas makna yang seragam.

Untuk menjaga stabilitas fitur teks, dilakukan juga filtering terhadap rare words, yaitu kata-kata yang muncul kurang dari dua kali di seluruh korpus data. Kata-kata ini dinilai tidak memiliki kontribusi yang signifikan terhadap proses klasifikasi dan justru dapat memperbesar risiko overfitting jika tetap dipertahankan. Dengan menghapus rare words, dimensi fitur menjadi lebih ramping dan representasi teks menjadi lebih stabil serta efisien secara komputasi.

Seluruh proses pembersihan teks ini kemudian dikemas ke dalam satu pipeline terpadu bernama full_clean_pipeline, yang memastikan setiap tahapan dijalankan secara sistematis dan berurutan. Hasil dari pipeline ini disimpan dalam kolom baru bernama cleaned_tweet, yang kemudian disinkronkan kembali dengan label multi-label hasil anotasi sebelumnya. Seluruh data yang telah diproses disatukan ke dalam satu dataframe dan disimpan dalam format CSV dengan nama scraping_labeled_tuned_cleaned_FIXED.csv, sebagai dataset akhir yang siap digunakan dalam proses modeling.

Untuk menangani ketidakseimbangan distribusi label, diterapkan strategi data augmentation yang difokuskan hanya pada label minor yaitu label yang memiliki jumlah kemunculan kurang dari 500 instance. Proses ini dilakukan dengan menyalin dan memodifikasi teks menggunakan struktur fungsi synonym_augment, yang pada tahap ini masih berupa dummy replacement, namun telah disiapkan untuk dikembangkan menggunakan thesaurus atau word embedding. Augmentasi dilakukan sebanyak tiga kali untuk setiap instance label minor, dengan tujuan memperbanyak variasi data tanpa harus menambah data manual secara eksplisit.

Tahapan terakhir adalah tokenisasi dan pembentukan kelas dataset, di mana setiap teks yang telah dibersihkan dikonversi menjadi representasi numerik menggunakan tokenizer dari model IndoBERT (indobenchmark/indobert-base-p1).

Tokenizer ini menghasilkan input_ids dan attention_mask yang diperlukan untuk pemrosesan teks oleh arsitektur transformer. Seluruh data kemudian dikemas dalam kelas khusus bernama CyberDataset, yang disiapkan untuk digunakan dalam DataLoader dari PyTorch, dengan pengaturan max_length=100, padding otomatis, serta label dalam format tensor float. Dataset kemudian dibagi menjadi data pelatihan dan validasi dengan rasio 80:20 agar performa model dapat diuji secara objektif terhadap data yang tidak pernah dilihat sebelumnya.

Dengan strategi data preparation yang menyeluruh dan disesuaikan dengan karakteristik media sosial, penelitian ini memastikan bahwa setiap aspek dari proses pembersihan, augmentasi, dan tokenisasi berjalan optimal. Kualitas data input yang bersih, konsisten, dan representatif menjadi landasan utama bagi model klasifikasi multi-label untuk mendeteksi berbagai bentuk kekerasan verbal dalam teks berbahasa Indonesia secara lebih akurat dan andal.

3.2.4 Modeling

Tahap modeling pada penelitian ini berfokus pada pengembangan sistem klasifikasi multi-label dengan menggunakan arsitektur *hybrid* yang menggabungkan IndoBERT dan Biderectional Long Short-Term Memory (BiLSTM). Pemilihan model ini didasarkan pada kelebihan IndoBERT sebagai model Transforme berbahasa Indonesia yang unggul dalam memahami konteks linguistik secara mendalam serta kemampuan dari BiLSTm dalam menangkap pola urutan kata dan dependensi temporal dari representasi yang dihasilkan oleh IndoBERT.

Model dirancang dengan memanfaatkan pre-trained IndoBERT yaitu *indobenchmark/indobert-base-p1* sebagai *feature extractor* untuk mengahasilkan representasi kontekstual dari setiap token. Hasil dari IndoBERT dimasukkan ke dalam lapisan BiLSTM yang mampu memproses informasi dalam dua arah untuk menangkap konteks secara menyeluruh. Output dari BiLSTm kemudian diproses melalui lapisan *dropout* dan dilanjutkan dengan lapisan *dense* sebagai *classifier* akhir dengan fungsi aktivasi *sigmoid* untuk mendukung prediksi multi-label.

Proses tokenisasi dilakukan menggunakan BertTokenizer dari IndoBERT dengan penyesuaian paarameter beruapa *padding* dan *truncation* agar setiap input memiliki panjang seragam dengan maksimal 100 token. Kemudian dataset dibagi menjadi dua bagian yaitu data pelatihan atau data train sebesar 80% dan data validasi atau data test sebesar 20%.

Proses pelatihan dijalankan selama beberapa *epoch* dengan menggunakan *optimizer* Adam dan fungsi loss *Focal Loss* yang dirancang utnuk memberikan penalti lebih besar terhadap kesalahan pada label yang sulit dikenali. Untuk menghasilkan keputusan atau hasil klasifikasi yang optimal, pengunaan metode *threshold* dinamis per label berbasis ROC-AUC bukan threshold tetap atau *default* seperti 0.5. Hal ini memungkinkan model memiliki ambang prediksi yang disesuaikan berdasarkan performa masing-masing label sehingga meningkatkan ketepatan dan akurasi prediksi multi-label secara keseluruhan.

3.2.5 Evaluation

Tahapa evaluasi dilakukan untuk mengukur performa model yang telah dilatih terhadap data validasi. Evaluasi dilakukan untuk setiap label secara terpisah serta dihitung juga skor rata-rata makro untuk memberikan gambaran menyeluruh terhadap performa model secara umum. Hal tersebut dilakukan karena klasifikasi pada penelitian ini bersifat multilabel.

Beberapa metrik evaluasi yang digunakan pada penelitian ini adalah *classification report, confusion matrix*, ROC AUC dan *threshold* berbasis ROC. *Classification report* digunakan untuk menampilkan metrik evaluasi berupa *precision, recall* dan *F1-score* untuk setiap label. *Precision* mengukur seberapa tepat model dalam memberikan label yang benar sedangkan *recall* menunjukkan sejauh mana model berhasil menemukan seluruh label yang relevan. Visualisasi distribusi prediksi model terhadap label aktual termasuk jumlah *true positive, false positive, false negative dan true negative* perlu dilakukan jika ingin megevaluasi kesalahan prediksi secara lebih rinci.

Metrik ROC AUC *(Receiver Operating Characteristic – Area Under Curve)* juga digunakan untuk menilai sejauh mana model mampu membedakan antara

kelas positif dan negatid pada setiap label. Metrik ini penting karena memberikan pandangan yang lebih objektif terhadap performa klasifikasi biner pada setiap label secara independen. Pada proses ini juga diterapkan threshold dinamis per label yang ditentukan berdasarkan kurva ROC.

3.2.6 Deployment

Tahap terakhir dalam metodologi CRISP-DM adalah *deployment*, yaitu proses implementasi model ke dalam bentuk aplikasi nyata yang dapat digunakan oleh pengguna akhir. Pada penelitian ini, model dikembangkan menjadi aplikasi web menggunakan platform Hugging Face Spaces dengan framework Gradio. Model klasifikasi yang telah dilatih kemudian dimuat ulang ke dalam antarmuka web interaktif memungkinkan pengguna untuk memasukkan teks secara langsung dan mendapatkan hasil prediksi secara otomatis berdasarkan 12 label kategori *cyberbullying*.

Pemilihan bentuk website untuk deployment dilakukan karena lebih praktis, universal, dan mudah diakses oleh siapa pun tanpa instalasi tambahan. Website dapat dijalankan langsung melalui browser dari berbagai perangkat (komputer, tablet, hingga smartphone), menjadikannya pilihan ideal untuk pengujian cepat dan demonstrasi publik. Selain itu, bentuk web memungkinkan evaluasi usability langsung dari sisi pengguna, serta mendukung distribusi hasil penelitian secara terbuka dan dapat direplikasi. Dibandingkan dengan *deployment* lokal atau berbasis API tertutup, aplikasi web bersifat lebih inklusif dan transparan, terutama dalam konteks penelitian akademik yang mendorong *reproducibility* dan aksesibilitas.

3.3 Teknik Pengumpulan Data

Teknik pengumpulan data dalam penelitian ini dilakukan melalui pendeketan *web scraping* yaiut proses pengambilan data secara otomatis dari platform media sosial Twitter. Data yang dikumpulkan berupa tweet atau teks berbahasa Indonesia yang berpotensi mengandung ujaran kebecian atau *cyberbullying.* Penelitian ini menggunakan alat bantu bernama Tweet Harvest yaitu sebuah script berbasis Python yang dirancang khusus untuk mengekstrak tweet berdasarkan kata kunci tertentu yang relevan dengan *cyberbullying.* Proses

scrapiong difokuskan pada tweet yang berbahsas Indonesia dengan kata-kata yang mengindikasikan potensi kekerasan verbal atau ujaran kebencian seperti kata ancaman , ejekan, kata kasar atau hinaan. Kata kunci yang digunakan telah disusun sebelumnya berdasarkan kategori-kategori cyberbullying, seperti penghinaan fisik, pelecehan gender, rasial, maupun ujaran kebencian terhadap kelompok tertentu.

3.4 Teknik Analisis Data

Teknis analisis data dalam penelitian ini dilakjukan melalu pendekatan berbasi data mining yang dirancang untuk mengolah data teks secara sistematis agar dapat diklasifikaikan ke dalam beberapa kategori *cyberbullying*. Proses analisis dimulai dari tahap pembersihan dan pra-pemrosesan data teks yang mencakup normaliasai kata , penghapusan simbol dan lainnya. Hasil dari preprocessing kemudian digunakan sebagai input untuk membangun dan melatih model klasigikasi multi-label.

Model utama yang digunakan dalam penelitian ini adalah model hybrid IndoBERT dan BiLSTm dimana IndoBERT berfungsi sebagai penghasil representasi kontekstual dari teks sedangkan BiLSTM digunakan untuk menangkap urutan dan pola kata melalui dua arah. Model dilatih menggunakan fungsi loss Focal Loss untuk mengataasi ketidakseimbangan kelas serta menggunakan *threshold* dinamis berbasis ROC-AUC untuk meningkatkan akurasi prediksi multi-label. Proses pelatihan dan validasi dilakukan dengan mmebagi dataset menjadi dua bagian yaitu 80% untuk data latih dan 20% untuk data validasi.

Peneliti melakukan evaluasi performa model dengan menggunakan beberapa me trik seperti *classification* report, confusion matrix dan ROC AUC per label setelah proses pelatihan selesai untuk melihat sejauh mana model mampu mengidentifikasi berbagi kategori *cyberbullying.* Model yang telah dilatih dan didebalusasi kemudian diimplementasikan aatu di deploy melalaui platform *Hugging Face Spaces* sehingga dapata digunakan untuku melakukan klasifikasi otomatis secara *real-time* terhadap teks baru. Proses analisis data tidak hanya sebatas berhenti pada validasi akademik saja tetapi juga diterapkan dalam bentuk sistem prediksi yang dapat diakses oleh pengguna lain.