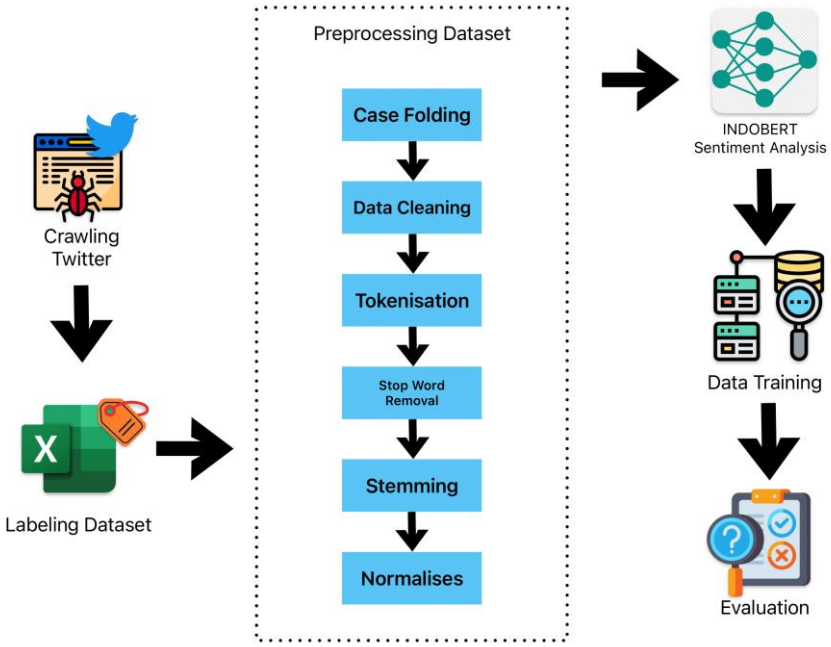
BAB III
METODOLOGI PENELITIAN

Pada bab ini akan membahas tentang perencanaan dalam analisis yang di gunakan dalam proses analysis sentimen kekerasan verbal terhadap data dari twitter dengan metode *Bidirectional Encoder Representations from Transformers* dengan model INDOBERT.

3.1 Arsitektur Umum

Metode yang diajukan dalam analisis sentimen terhadap kekerasan verbal di twitter dengan metode *Bidirectional Encoder Representations from Transformers* terdiri daribeberapa langkah seperti yang tertera pada Gambar 3.1.



Gambar 3. 1 Langkah Penelitian

Struktur dari sistem di atas mengindikasikan bahwa langkah awal dalam analisis sentimen adalah melakukan Crawling atau ekstraksi data dari platform Twitter, yang akan berfungsi sebagai sumber data. Hasil dari proses Crawling ini kemudian dikompilasi menjadi sebuah dataset. Setelah itu, dataset mendapatkan anotasi berupa label negatif, netral, atau positif. Dataset yang telah dlabeling ini kemudian menjalani tahap preprocessing. Proses preprocessing ini bertujuan untuk mengorganisir data yang sebelumnya tidak teratur menjadi data yang lebih terstruktur melalui serangkaian langkah, termasuk case folding, pembersihan data, tokenisasi, penghapusan kata-kata tak berarti (stopwords), stemming, dan normalisasi untuk Bahasa yang kurang baku. Setelah melalui rangkaian langkah tersebut, dataset yang telah melalui proses preprocessing diarahkan untuk dilatih agar dapat dikelompokkan menjadi tiga kategori: negatif, netral, dan positif, menggunakan metode INDOBERT. Selanjutnya, hasil klasifikasi dievaluasi untuk melihat hasil akhirnya.

3.2 Data Crawling

Pengambilan data menggunakan Twint memiliki kelebihan yaitu dapat menghindari batasan pengambilan data yang diberlakukan oleh Twitter, serta dapat mengambil data secara terstruktur sehingga memudahkan dalam analisis data. Namun, penggunaan Twint juga memiliki beberapa kekurangan seperti keterbatasan dalam mengambil data yang bersifat privasi atau terbatas, serta kurangnya validasi data yang diambil. Oleh karena itu, peneliti perlu mempertimbangkan kelebihan dan kekurangan dari penggunaan Twint dalam mengambil data pada penelitiannya.

Proses pengambilan data Twitter dibantu dengan bahasa pemrograman *phyton.* Proses pengambilan data *tweet* akan dilakukan filtering dari tanggal 10 Februari 2023 sampai 13 Februari 2023. Pada tabel 2 berikut adalah kategori kekerasan verbal yang digunakan sebagai acuan Crawler untuk mengambil data pada twitter :

Tabel 2 Kategori Kekerasan Verbal

Kategori	Keterangan	Keyword
Bentuk Kekerasan Verbal Mengumpat	Mengeluarkan perkataan yang buruk kepada seseorang.	- Sampah - Lonte
Bentuk Kekerasan Verbal Eufemisme	Eufemisme merupakan pengucapan gaya bahasa halus untuk menyindir atau mengkritik dengan nada yang terkesan melecehkan.	- Cacat mental - Penjilat - Belajar lagi
Bentuk Kekerasan Verbal Disfemisme	Disfemisme adalah perkataan dengan konotasi yang kasar, negatif, tidak sopan, menyinggung, dan menyakiti perasaan orang lain untuk memperlihatkan sikap ketidak sukaan, kemarahan, dan rasa benci dengan mengkasarkan, mengeraskan fakta melalui ucapan sehingga maknanya berbeda dari sebenarnya.	- Goblok - Tolol - Bodoh
Kekerasan Verbal Stigmatisasi	Stigmatisasi pemberian “tanda‟ atau stigma terhadap seseorang atau sekelompok orang dengan pengertian yang bermakna tertentu dalam situasi dan konteks tertentu secara terbuka atau terselubung untuk mempengaruhi daya pikir atau daya evaluasi seseorang atau sekelompok orang terhadap sesuatu, demi kepentingan si pemberi stigma	- Kafir - Murahan - Dajjal
Kekerasan Verbal Hiperbola	Hiperbola adalah ungkapan yang dilebih-lebihkan sehingga tidak sesuai dengan sebenarnya	- Kebal hukum - Tidak punya otak
Bentuk Kekerasan Verbal Asosiasi pada Binatang	Mengibaratkan atau menyamakan seseorang secara negatif pada binatang atau perkataan negative yang diberikan kepada orang lain, tetapi berasosiasi pada binatang.	- Babi - Anjing

3.3 Labeling Data

3.3.1 Labeling Data Manual

Proses Labeling data secara manual dilakukan sebagai langkah awal dalam persiapan dataset berlabel untuk analisis sentimen. Dataset yang telah di-sampling dari seluruh dataset akan dilabeli oleh tiga anotator berbeda. Anotator akan memberikan label pada setiap data berdasarkan sentimen yang terkandung dalam teks-tweet. Ketentuan label yang digunakan adalah angka "2" untuk sentimen negatif, angka "1" untuk sentimen positif, dan angka "0" untuk sentimen netral. Penggunaan label angka ini dipilih untuk mempermudah dan mempercepat proses annotator dalam melabeli data, sehingga sentimen pada setiap data dapat diidentifikasi dengan lebih cepat dan tepat.

Proses Labeling data secara manual ini menjadi langkah krusial dalam persiapan dataset berlabel yang akan menjadi dasar bagi model dalam proses Semi-Supervised Learning. Dengan adanya ketentuan label yang jelas dan efisiensi dalam proses labeling, diharapkan analisis sentimen pada data berbahasa Indonesia dapat dilakukan secara lebih efektif dan akurat. Dataset yang telah dilabeli oleh anotator akan menjadi dataset latih yang berharga untuk mengajarkan model dalam mengenali sentimen dalam teks-tweet berbahasa Indonesia. Selanjutnya, dataset berlabel ini akan digunakan dalam eksperimen Semi-Supervised Learning untuk mengoptimalkan performa model analisis sentimen.

3.3.2 Labeling Data Otomatis

Dalam metode supervised learning untuk analisis sentimen, diperlukan kumpulan data yang telah diberi label atau anotasi oleh annotator. Labelisasi ini penting karena metode supervised learning memerlukan contoh untuk melakukan

generalisasi dan prediksi. Artinya, model dapat melihat contoh dan menghasilkan keluaran yang sesuai dengan label yang diinginkan (Goldberg, 2017). Model tersebut dapat memahami komentar dengan sentimen negatif, netral, atau positif. Labelisasi dilakukan dengan tujuan untuk menentukan kategori sentimen dari setiap komentar, yaitu negatif, netral, atau positif, dengan memberikan nilai sebagai tanda. Komentar dengan sentimen positif diberi nilai 2, sementara komentar dengan sentimen netral diberi nilai 1 dan komentar dengan sentimen negatif diberi nilai 0. Contoh dataset yang sudah dilabelisasi sebagaimana terlihat pada Tabel 3.

Tabel 3 Contoh Dataset

Komentar	Sentimen
Nyuruh mak nya buat mandi lumpur, emang gk ada otak anak ini.	2
Minimal mandi dulu mbak, biar keliatan terang	1
Gk punya bapak ya?	1
Mendingan pak john nawarin kerja ke saya aja pak.	0
Ngoceh terus lu njing	2
Muka berbie kelakuan kaya iblis	2
Baik sekali pemuda ini, mau mengingatkan temannya untuk jangan lupa makan siang, walaupun lagi bulan puasa.	0
Kaya minta duit tebusan	1

Model supervised learning dibangun menggunakan Sequential API dari TensorFlow dengan menggunakan layer Embedding untuk menerjemahkan kata-kata menjadi vektor numerik, serta dua layer LSTM dengan arah maju dan mundur (Bi- LSTM) untuk mengambil konteks dari teks secara lebih lengkap. Terakhir, dilakukan layer Dense dengan fungsi aktivasi softmax yang menghasilkan tiga output untuk klasifikasi sentiment, agar bisa menghasilkan 3

jenis label yaitu label 0 untuk netral, label 1 untuk positif dan label 2 untuk negatif.

Pada bagian learning rate, dilakukan penjadwalan agar learning rate dapat menyesuaikan diri selama proses pelatihan. Kemudian, model dikompilasi menggunakan fungsi Adam sebagai optimizer dengan sparse categorical crossentropy sebagai fungsi kerugian (loss) untuk tugas klasifikasi multi-kelas. Metrik sparse categorical accuracy digunakan untuk mengukur akurasi pada model.

3.4 Pre-Processing Data

Dalam penelitian ini, dilakukan proses preprocessing untuk mengubah dataset yang memiliki struktur yang kurang terorganisir menjadi lebih terstruktur, sehingga mempermudah pengolahan data dengan melibatkan langkah-langkah seperti case folding, data cleaning, tokenisasi, stopwords removal, stemming, dan normalisasi. Selain itu, melalui proses preprocessing ini, hasil analisis sentimen dapat diharapkan menjadi lebih optimal.

Proses Preprocessing akan dilakukan pada beberapa tahap yang berbeda, dimulai dengan sebelum tahap labeling data secara manual untuk membersihkan data dari karakter yang tidak relevan, sehingga memudahkan *annotator* dalam membaca data dan memberikan sentimen, yang kedua pada saat data akan di labeling secara otomatis dengan supervised-learning agar data dapat lebih seragam dan lebih baik untuk melatih model Supervised Learning. Beberapa tahap Preprocessing yang akan di gunakan adalah sebagai berikut :

3.4.1 Case Folding

Case folding dilakukan dengan mengubah semua huruf kapital (uppercase)

dalam dataset menjadi huruf kecil (lowercase). Langkah ini diimplementasikan untuk menyamakan semua karakter dalam dataset menjadi format huruf kecil. Transformasi ini berguna untuk memperoleh generalisasi yang lebih baik sehingga kata "Saya" dan "saya" akan dianggap setara. Proses case folding menggunakan fungsi lower() yang disediakan oleh library Python. Ilustrasi kasus case folding dapat dilihat pada Tabel 4.

Tabel 4 Perbandingan hasil Case Folding

Komentar	Hasil *Case Folding*
NGGAK ADA OTAK!!	nggak ada otak

3.4.2 Data Cleaning

Pada langkah ini, frasa-frasa dalam dataset dijernihkan dari semua elemen yang mungkin mempengaruhi hasil analisis, seperti kata-kata yang memiliki karakter berulang lebih dari satu kali, tautan, nama pengguna (*@username*), tanda pagar (#), angka, simbol-simbol, kelebihan spasi, tanda baca, dan bilangan. Proses pembersihan data dilakukan dengan menggunakan ekspresi regular untuk pencocokan dan penghapusan. Ilustrasi langkah pembersihan data dapat ditemukan pada Tabel 5 sebagai contohnya.

Tabel 5 Perbandingan hasil Data Cleaning

Komentar	Hasil *Data Cleaning*
@goodbabyborn Goodbye Logika 🔥🔥🔥!!!!!!!	Goodbye logika

3.4.3 Tokenisasi

Tokenisasi merupakan tahap di mana kalimat-kalimat dibagi menjadi fragmen kata, tanda baca, dan ekspresi yang memiliki makna sesuai dengan norma bahasa yang digunakan. Dalam tahap ini, penulis menggunakan fungsi

word_tokenize yang disajikan oleh perpustakaan NLTK. Ilustrasi dari langkah tokenisasi dapat ditemukan dalam Tabel 6 sebagai contoh.

Tabel 6 Perbandingan hasil Tokenisasi

Komentar	Hasil Tokenisasi
Muka berbie kelakuan kaya iblis	‘muka’ ‘barbie’ ‘kelakuan’ ‘kaya’ ‘iblis’

3.4.4 Stopwords Removal

Proses Stopwords Removal adalah tahapan penting dalam analisis sentimen untuk menghapus kata-kata yang tidak memiliki makna atau relevansi dalam konteks analisis. Pada tahap ini, penuilis menggunakan library stopwords Bahasa Indonesia yang disediakan oleh indoNLP, yang mencakup daftar kata-kata umum yang dianggap sebagai stopwords [17]. Contoh tahap stopwords removal yang dapat dilihat pada Tabel 7.

Tabel 7 Perbandingan hasil Stopwords Removal

Komentar	Hasil Tokenisasi
Bukannya nolongin malah ngerekam juga si anjing	‘bukan’ ’nya’ ‘nolongin’ ‘malah’ ‘ngerekam’ ‘si’ ‘anjing’

3.4.5 Stemming

Stemming adalah proses yang dilakukan untuk mengubah kata yang memiliki imbuhan menjadi kata dasarnya (*root form*) dengan menghapus imbuhan seperti prefiks, sufiks, dan konfiks. Pada tahap ini, *stemming* dilakukan dengan menggunakan *library* Sastrawi. Contoh tahap stemming yang dapat dilihat pada Tabel 8.

Tabel 8 Perbandingan hasil Stemming

Komentar	Hasil *Stemming*

“ Mendingan beli skripsi ”	‘“ Mending beli skripsi ”

3.4.6 Normalisasi

Tahap normalisasi merupakan langkah krusial dalam analisis sentimen, di mana dataset yang mengandung kata-kata tidak baku atau tidak standar akan diubah menjadi kata-kata standar atau sesuai dengan ejaan yang benar. Dalam tahap ini, penulis menggunakan metode penghapusan slangword dari indonlp untuk menangani berbagai kata slang seperti "bgt", "trus", "slalu", "pngn", "aq", "kereeen", "kereenn", dan lain-lain. Tanpa normalisasi, sistem akan memperlakukan kata-kata tersebut sebagai entitas yang berbeda, padahal sebenarnya memiliki makna yang serupa, seperti dalam contoh kata "keren". Oleh karena itu, normalisasi diterapkan untuk memastikan bahwa data yang diolah sudah berada dalam bentuk standar dan sesuai dengan aturan ejaan yang berlaku. Dengan menggunakan metode replace_slangword dari indoNLP[17], penuilis berharap dapat meningkatkan kualitas analisis sentimen dengan mengonversi kata-kata slang menjadi bentuk yang benar dan lebih mudah dipahami.Proses normalisasi ini menggunakan Kamus Alay2 [18]. Sebagai tambahan, kata tidak baku yang diperoleh saat menganotasi juga ditambahkan. Contoh tahap normalisasi yang dapat dilihat pada Tabel 9.

Tabel 9 Perbandingan hasil Normalisasi

Komentar	Hasil Normalisasi
Hajar bang , abisin semua giginya tanpa sisa	‘hajar’ ‘bang’ ‘habis’ ‘semua’ ‘gigi’ ‘tanpa’ ‘sisa’

3.5 Impementasi INDOBERT

Dalam rangka penelitian ini, peneliti memilih untuk mengadopsi model INDOBERT sebagai kerangka kerja analisis. Keputusan ini didasarkan pada kenyataan bahwa INDOBERT telah menjalani pelatihan pada lebih dari 220 juta kata Bahasa Indonesia. Setelah tahap pengolahan awal data, dataset akan diubah menjadi format yang dapat diterima oleh IndoBERT, yaitu dalam bentuk vektor representasi kata menggunakan alat bantu IndoBERT Tokenizer. Langkah selanjutnya melibatkan tahapan fine-tuning terhadap *hyper-parameter*, di mana model INDOBERT yang telah dilatih sebelumnya disesuaikan untuk melakukan tugas klasifikasi sentimen.

3.6 Evaluasi

Tahapan evaluasi ditujukan untuk melihat hasil labeling secara supervised learning dan analisis sentimen terhadap kalimat yang ada pada dataset. Pada analisis sentimen kekerasan verbal dengan model INDOBERT, confusion matrix dapat digunakan untuk mengevaluasi kinerja model dalam melakukan klasifikasi komentar sebagai positif, netral, atau negative. Nilai akurasi tertinggi yang didapatkan dari proses training sebelumnya akan menjadi nilai akurasi model. Untuk memperoleh prediksi dari model, confusion matrix digunakan sebagaimana terlihat pada Tabel 10.

Tabel 10 Confussion Matrix

		True Class		
		Positive	Netral	Negative
Predicted Class	Positive	True Positive (TP)	False Positive (FNt)	False Positive (FP)
	Neutral	False Neutral (FNt)	True Neutral (TNt)	False Neutral (FNt)
	Negative	False Negative (FN)	False Negative (FNt)	True Negative (TN)

Kategori pada *confusion matrix* terdiri dari empat kategori, yaitu *True Positive* (TP), *False Positive* (FP), *True Neutral* (TNt), *False Neutral* (FNt), *True Negative* (TN), dan *False Negative* (FN). *True Positive* adalah kalimat memiliki sentimen positif dan hasil prediksinya juga menunjukkan sentimen positif. *False Positive* (FP) adalah kalimat yang memiliki sentimen positif tetapi hasil prediksinya menunjukkan sentimen netral atau negatif. *True Neutra*l (TNt) adalah kalimat memiliki sentimen netral dan hasil prediksinya juga menunjukkan sentimen netral. *False Neutral* (FNt) adalah kalimat yang memiliki sentimen netral tetapi hasil prediksinya menunjukkan sentimen positif atau negatif. *True Negative* (TN) adalah kalimat yang memiliki sentimen negatif dan hasil prediksinya juga menunjukkan hasil negatif. *False Negative* (FN) adalah kalimat yang memiliki sentimen negatif tetapi hasil prediksinya menunjukkan sentimen netral atau positif.

Setelah mendapatkan nilai untuk *confusion matrix*, nilai *accuracy, precision, recall,*dan *F-measure* juga dapat diperoleh. *Accuracy* bertujuan untuk menunjukkan persentasi dari input yang berhasil diprediksi oleh neural network dengan benar. Nilai akurasi akan semakin baik ketika nilai *loss* menurun. *Precision* bertujuan untuk menghitung persentase dari input yang dideteksi oleh sistem misalnya, sistem memberi label input sebagai positif yang pada aslinya adalah positif juga. *Recall* bertujuan untuk menghitung persentase dari input yang diidentifikasi True secara benar oleh sistem. Sedangkan F-*measure* adalah rata-rata yang diperoleh dari *precision* dan *recall*. Rumus perhitungan untuk mendapatkan *accuracy* dan *F-score* ditunjukkan pada persamaan 3.1 dan persamaan 3.4. Sedangkan rumus untuk perhitungan *precision* dan *recall* dilakukan untuk tiap sentimen dengan contoh seperti pada persamaan 3.2 dan 3.3.

$$Accuracy = \frac{TP+Tnt+TN}{TP+FP+TNt+FNt+TN+FN} \quad \text{.....(3.1)}$$

$$Precision\ Positive = \frac{TP}{TP+FP} \quad \text{.....(3.2)}$$

$$Recall\ Positive = \frac{TP}{TP+FN+FNt} \quad \text{.....(3.3)}$$

$$F - measure = 2 * \frac{Precision*Recall}{Precision+Recall} \quad \text{.....(3.4)}$$