BAB II

TEORI PENELITIAN

2.1 Related Works

Penelitian mengenai deteksi ujaran kebencian dan analisis sentimen pada media sosial telah banyak dilakukan dengan pendekatan yang beragam, baik menggunakan metode *machine learning* tradisional maupun model *deep learning* terkini.berikut merangkum beberapa penelitian terdahulu yang relevan dengan fokus pada metode yang digunakan dan hasil yang diperoleh dalam klasifikasi teks bermuatan negatif atau ofensif:

Tabel 2.1 Daftar Penelitian Terdahulu

Penulis	Judul	Jurnal	Metode	Hasil
Dhineshkumar P. & Dr. A. Nithya (2024) [16]	Hate Speech *Classification* Using RFDT, BiLSTM, and BiLSTM	Journal of Propulsion Technology (Vol. 45 No. 1)	RFDT, BiLSTM, RFDT+BiLSTM	**RFDT**: - Akurasi 76.45%, - Recall 72.23%, - Precision 74.46%; **BiLSTM**: - Akurasi 69.23%, - Recall 68.26%, - Precision 67.17%; **RFDT+BiLSTM**: - Akurasi 73.68%, - Recall 70.49%, - Precision 71.53%
R. A. Saputra and Y. Sibaroni [17]	A Multilabel Hate Speech Classification in Indonesian Political Discourse on X using Combined Deep Learning Models with Considering Sentence Length Text Detection	Jurnal Ilmu Komputer dan Informasi (JIKI)	BERT-BiLSTM	**Akurasi** sebesar 83.10%
Ayu Fransiska et al. (2022) [18]	Algoritme Logistic Regression untuk Mendeteksi Ujaran	Jurnal Nasional Komputasi dan Teknologi Informasi, Vol. 5 No. 4	Logistic Regression	Akurasi rata-rata: 75,59%; Kelas HateSpeech: 75,86%, Abusive: 80,05%, Level: 70,86% menggunakan fitur seleksi dan optimasi parameter C & solver

Penulis	Judul	Jurnal	Metode	Hasil
	Kebencian dan Bahasa Kasar Multilabel pada Twitter Berbahasa Indonesia			
Leno Dwi Cahya et al. (2023) [19]	Improving Multi-label *Classification* Performance on Imbalanced Datasets Through SMOTE Technique and Data Augmentation Using IndoBERT Model	Jurnal Nasional Teknologi dan Sistem Informasi, Vol. 9 No. 3	IndoBERT + SMOTE + Augmentation	SMOTE: Akurasi 82%, Precision 87%, Recall 85%, F1-score 86%; Augmentasi: Akurasi 78%, Precision 85%, Recall 82%, F1-score 83%
Fauzi Ihsan et al. (2021) [20]	Algoritme Decision Tree untuk Mendeteksi Ujaran Kebencian dan Bahasa Kasar Multilabel pada Twitter Berbahasa Indonesia	Jurnal Teknologi dan Sistem Komputer, Vol. 9 No. 4	Decision Tree dengan rekayasa fitur (leksikon, tekstual, khusus)	Akurasi tertinggi dengan fitur leksikon: 70,48% (90:10) dan 69,54% (80:20); Rata-rata presisi & sensitivitas juga meningkat pada kelas Abusive dan HateSpeech
Imamah Nur Fadlilah et al. (2025) [21]	Komparasi Metode Label Powerset K-NN dan ML-KNN dalam Klasifikasi Multi-Label *Cyberbullying* pada Komentar Instagram	JATI (Jurnal Mahasiswa Teknik Informatika), Vol. 9 No. 3	Label Powerset KNN dan ML-KNN dengan TF-IDF (Unigram, Bigram, Trigram)	Model terbaik: ML-KNN dengan TF-IDF Unigram, Akurasi: 64%, F1-Score: 0.74, Hamming Loss: 0.13; ML-KNN mengungguli LP-KNN di semua metrik evaluasi
Elita Aurora Az Zahra et al. (2023) [22]	*Classification* of Multi-Label of Hate Speech on Twitter Indonesia using LSTM and BiLSTM Method	JINAV: Journal of Information and Visualization, Vol. 4 No. 2	LSTM dan BiLSTM	LSTM: Akurasi 78.67% (epoch 10), BiLSTM: Akurasi 80.25% (epoch 10); BiLSTM konsisten lebih unggul daripada LSTM dalam klasifikasi multi-label hate speech

Penulis	Judul	Jurnal	Metode	Hasil
Ari Muzakir et al. (2022) [23]	A Comparative Analysis of *Classification* Algorithms for *Cyberbullying* Crime Detection	Scientific Journal of Informatics, Vol. 9 No. 2	Naïve Bayes, Decision Tree, Logistic Regression, SVM + Bag of Words (Unigram, Bigram, Trigram)	SVM unggul dengan akurasi 76% dan F1-score 82%; DT memiliki recall tertinggi 84%; BOW kombinasi fitur memberi hasil optimal untuk SVM
Sahinur Rahman-Laskar et al. (2024) [24]	*Cyberbullying* Detection in a Multi-*classification* Codemixed Dataset	Computación y Sistemas, Vol. 28 No. 3	SVM, Random Forest, Logistic Regression, Naive Bayes, XGBoost, Ensemble (Voting)	Model terbaik: Ensemble Voting Classifier dengan Akurasi: 60.9%, F1-Score: 0.5879; XGBoost unggul pada kelas tertentu (F1-score Age: 0.77, Gender: 0.71, Abusive: 0.55)
Fajar Agus Maulana & Iin Ernawati (2020) [25]	Analisa Sentimen *Cyberbullying*di Jejaring Sosial Twitter dengan Algoritma Naïve Bayes	Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)	Naïve Bayes	Akurasi: 76%, Precision: 76,09%, Recall: 97,22%, Specificity: 21,4%

Berdasarkan Tabel 2.1, berbagai penelitian terdahulu telah mengeksplorasi metode untuk mendeteksi dan mengklasifikasikan *cyberbullying* dalam teks baik melalui pendekatan *machine learning* konvensional maupun model *deep learning*. Penelitian yang dilakukan [16] menggunakan RFDT dan BiLSTM berhasil mencapai akurasi hingga 73,68% untuk klasifikasi ujaran kebencian multilabel. Meskipun demikian, penelitian ini terbatas pada data berbahasa Inggris dan tidak menyentuh konteks linguistik khas Indonesia seperti bahasa alay, singkatan atau istilah lokal yang kerap muncul di media sosial. Hal ini menjadi celah penting yang coba diisi oleh penelitian ini dengan menerapkan BiLSTM pada data Twitter berbahasa Indonesia.

Penelitian [22] secara eksplisit membandingkan performa LSTM dan BiLSTM pada data ujaran kebencian di Twitter Indonesia dan menemukan bahwa BiLSTM lebih unggul. Namun, penelitian ini terbatas pada metrik akurasi dan tidak mengoptimalkan performa melalui teknik *augmentasi* atau penanganan

ketidakseimbangan data yang menjadi sebuah aspek penting yang diadopsi dalam penelitian ini dengan pendekatan oversampling dan *Focal Loss*.

Kritik juga muncul dari penelitian [20] yang masih menggunakan algoritma klasik seperti *Decision Tree* dan *Logistic Regression*. Walaupun memiliki nilai akurasi cukup baik, pendekatan tersebut cenderung kurang optimal dalam menangkap konteks semantik kata secara berurutan yang justru menjadi kekuatan utama dari arsitektur BiLSTM. Fitur *leksikal* yang digunakan pada penelitian mereka bergantung pada pembuatan manual dan kamus yang terbatas sehingga rentan terhadap perubahan gaya bahasa di media sosial.

Penelitian [19] menggaris bawahi pentingnya augmentasi data dan penyeimbangan kelas, namun metode mereka masih terfokus pada model IndoBERT yang sangat bergantung pada sumber daya pemrosesan tinggi dan tidak dibuktikan kinerjanya dalam konteks multilabel *cyberbullying* secara spesifik. Sebaliknya, penelitian ini memanfaatkan BiLSTM yang lebih ringan namun tetap mampu mempelajari hubungan kata dua arah serta menerapkan teknik augmentasi dan balancing agar dapat menangani label-label minoritas dengan lebih baik.

Adapun studi [24] yang menggunakan pendekatan ensemble dan model klasik namun belum menyentuh persoalan multilabel imbalance dan cenderung hanya mengandalkan metrik akurasi tanpa evaluasi makro seperti *F1-score* atau *ROC-AUC*. Oleh karena itu metode evaluasi yang lebih menyeluruh dan spesifik terhadap konteks *cyberbullying* diterapkan dalam penelitian ini.

Dengan memperhatikan berbagai kekurangan dan keterbatasan dari penelitian-penelitian sebelumnya, penelitian ini menjadi penting untuk dilakukan. Fokus utamanya adalah membangun model multilabel *classification* berbasis BiLSTM untuk mendeteksi berbagai jenis *cyberbullying* pada teks berbahasa Indonesia dari media sosial Twitter dan Instagram secara lebih akurat, efisien dan adaptif terhadap dinamika bahasa digital.

Penelitian-penelitian terdahulu yang disebutkan memiliki keterbatasan dalam menangani data berbahasa Indonesia dengan konteks linguistik yang khas, seperti bahasa alay, singkatan, dan istilah lokal. Selain itu, beberapa studi hanya

berfokus pada metrik akurasi tanpa mengoptimalkan performa melalui teknik augmentasi atau penanganan ketidakseimbangan data secara komprehensif. Penggunaan algoritma klasik juga menunjukkan keterbatasan dalam menangkap konteks semantik kata secara berurutan dan bergantung pada fitur leksikal manual yang terbatas.

Penelitian ini mengatasi celah tersebut dengan menggunakan pendekatan hibrida IndoBERT dengan BiLSTM pada data Twitter berbahasa Indonesia, yang secara eksplisit disebutkan dalam Bab 4 sebagai peningkatan. IndoBERT dipilih karena kemampuannya memahami konteks linguistik bahasa Indonesia secara mendalam, sementara BiLSTM digunakan untuk menangkap urutan kata secara dua arah, yang diharapkan mampu mengklasifikasikan berbagai bentuk cyberbullying dalam skema multi-label. Selain itu, penelitian ini mengadopsi teknik augmentasi dan Focal Loss untuk menangani ketidakseimbangan data secara efektif , serta menerapkan evaluasi makro seperti F1-score dan ROC-AUC. Peningkatan ini secara rinci dianalisis dan dibahas pada Bab 4, khususnya pada bagian "Analisis dan Hasil Penelitian".

2.2 Teori Penelitian

2.2.1 Analisis Sentimen

Dari perspektif psikologi, sentimen merupakan bentuk sikap atau kecenderungan psikologis yang muncul sebagai hasil dari interaksi antara emosi, perasaan dan suasana hati (*mood*) sesemanusia terhadap suatu objek tertentu [26]. Objek ini dapat berupa individu, benda, peristiwa, ide atau fenomena sosial yang menjadi pusat perhatian seseorang. Sentimen tersebut dapat bersifat positif, negatif atau netral tergantung pada bagaimana individu menafsirkan dan merespons pengalaman atau stimulus tertentu yang diterimanya [27]. Tiga komponen penting yang membentuk dan mempengaruhi sentimen ini termasuk emosi, perasaan dan mood, sejatinya merupakan aspek yang saling berkaitan dalam studi psikologi [28], [29].

Emosi adalah reaksi kompleks yang mencakup komponen psikologis dan fisiologis sebagai respons terhadap stimulus eksternal atau internal [28]. Ketika

emosi muncul, tubuh dan pikiran individu merespons dalam bentuk ekspresi wajah, perubahan nada suara maupun pola pikir tertentu. Perasaan dalam hal ini merupakan manifestasi sadar dari emosi yang dialami yaitu ketika individu mampu mengenali dan mengartikulasikan bahwa ia sedang sedih, senang, marah atau takut. Sementara itu, mood atau suasana hati adalah keadaan emosional yang berlangsung lebih lama dan sering kali tidak memiliki pemicu yang jelas [28]. Mood dapat memengaruhi kecenderungan sentimen sesemanusia dalam jangka waktu yang lebih panjang dan secara signifikan berperan dalam cara individu membentuk opini atau penilaian terhadap sesuatu.

Ketiga aspek psikologis ini mempengaruhi sentimen sesemanusia karena mereka mengonstruksi cara pandang dan interpretasi terhadap informasi. Sesemanusia yang sedang berada dalam suasana hati bahagia, cenderung merespons berbagai hal secara positif. Individu yang sedang marah atau tertekan akan lebih mudah menilai hal-hal secara negatif. Dalam kondisi netral, sentimen yang ditunjukkan pun cenderung datar dan tidak menunjukkan kecenderungan tertentu terhadap objek yang diamati.

Dalam ranah komputasi, konsep sentimen ini diadopsi dalam bidang yang disebut analisis sentimen yaitu proses sistematis untuk mengidentifikasi, mengklasifikasi dan mengukur nada emosional yang terkandung dalam teks digital, seperti komentar, ulasan atau unggahan di media sosial. Tujuan utama dari analisis sentimen adalah untuk menentukan apakah teks tersebut mengandung muatan emosi yang bersifat positif, negatif atau netral [30]. Teknik ini menjadi sangat relevan terutama dalam konteks pemantauan opini publik, pengukuran reputasi merek, analisis perilaku konsumen hingga pengawasan terhadap ujaran kebencian di internet.

Analisis sentimen dilakukan dengan pendekatan *natural language processing* (NLP) dan *machine learning* yang memungkinkan sistem untuk memproses bahasa alami, mengekstraksi fitur linguistik dan mengklasifikasikan teks berdasarkan tema atau aspek tertentu. Terdapat berbagai jenis pendekatan dalam analisis sentimen termasuk analisis sentimen berbasis aspek (*aspect-based*

sentiment analysis) [31], deteksi emosi (*emotion detection*) [32] dan pendekatan granular (*fine-grained*) yang memberikan tingkat kehalusan analisis dari polaritas yang sederhana hingga kompleks [33]. Analisis sentimen banyak digunakan untuk pemantauan media sosial, pengelolaan layanan pelanggan serta sebagai alat pendukung dalam penelitian pasar dan kebijakan publik.

2.2.2 Platform Twitter / X

Twitter yang kini secara resmi dikenal sebagai X, merupakan salah satu platform media sosial paling populer dan berpengaruh dalam ekosistem digital global. Didirikan pada tahun 2006 oleh Jack Dorsey dan diluncurkan pada bulan Juli tahun yang sama. Twitter mulanya dikenal sebagai layanan *microblogging* yang memungkinkan pengguna membagikan pemikiran, opini dan informasi dalam bentuk tweet dengan batasan 140 karakter yang kemudian ditingkatkan menjadi 280 karakter [34]. Seiring waktu, platform ini telah berevolusi menjadi ruang diskursus publik utama yang digunakan oleh individu, tokoh publik, jurnalis, pemerintah dan organisasi di seluruh dunia. *Rebranding* menjadi X pada pertengahan 2023 oleh Elon Musk menandai transformasi platform ini menjadi *super app*, namun *core* fungsinya sebagai platform opini terbuka tetap dipertahankan.

Twitter memiliki struktur komunikasi terbuka yang unik bagi interaksi membahas topik apa pun. Setiap pengguna dapat menyebut (*mention*) akun lain, membalas (*reply*), menyebarluaskan informasi melalui fitur retweet dan membentuk opini publik melalui *trending topics* serta penggunaan *hashtag*. Keterbukaan sistem ini menjadikan Twitter sebagai medan yang sangat dinamis sekaligus rentan terhadap penyalahgunaan termasuk dalam bentuk *cyberbullying*. Kecepatan informasi menyebar, minimnya moderasi terhadap ujaran tertentu serta sifat anonimitas pengguna berkontribusi pada tingginya prevalensi kekerasan verbal dalam platform ini [35].

Menurut laporan dari Statista (2025), Twitter/X memiliki lebih dari 586 juta pengguna aktif bulanan di seluruh dunia, dengan Indonesia menjadi salah satu negara dengan pengguna aktif terbanyak di kawasan Asia Tenggara [36]. Studi dari Amnesty International menunjukkan bahwa Twitter menjadi salah satu media sosial

dengan tingkat tertinggi dalam hal penyebaran ujaran kebencian dan *cyberbullying* terutama terhadap kelompok minoritas dan perempuan. Di Indonesia, laporan dari Kominfo (2023) juga mencatat bahwa sebagian besar aduan masyarakat terkait konten negatif di media sosial berasal dari Twitter termasuk di dalamnya penghinaan, pelecehan serta ancaman berbasis identitas [37].

Dalam dunia riset dan pengembangan sistem deteksi teks otomatis, Twitter/X menjadi sumber data yang sangat berharga. Platform ini menyediakan Twitter API (*Application Programming Interface*) yang memungkinkan peneliti dan pengembang untuk melakukan scraping data berupa tweet, metadata pengguna, interaksi sosial dan lain sebagainya. Versi terbaru dari Twitter API v2 menyediakan endpoint yang mendukung pencarian tweet berdasarkan kata kunci, waktu, lokasi, bahasa serta parameter konteks lainnya [38]. Dengan penggunaan autentikasi berbasis token dan OAuth 2.0, API ini memungkinkan pengumpulan data dalam skala besar secara terstruktur dan legal meskipun tetap dibatasi oleh kebijakan rate-limit yang diberlakukan oleh pihak platform.

Melalui pemanfaatan Twitter API, peneliti dapat membangun korpus data berbahasa Indonesia yang relevan dengan studi *cyberbullying* termasuk dalam mengklasifikasikan bentuk-bentuk ujaran seperti abusive, rasis, *body shaming*. Oleh karena itu, Twitter/X bukan hanya menjadi medan terjadinya fenomena *cyberbullying*, tetapi juga sumber utama yang memungkinkan dikembangkannya sistem kecerdasan buatan untuk mengidentifikasi dan menanggulangi bentuk kekerasan digital tersebut secara otomatis dan *real time*.

2.2.3 *Cyberbullying*

Cyberbullying merupakan bentuk kekerasan berbasis teknologi yang dilakukan melalui perangkat digital seperti ponsel, komputer dan tablet dengan menggunakan platform media sosial, pesan teks, aplikasi chatting, forum daring dan bentuk komunikasi digital lainnya [39]. Fenomena ini berkembang pesat seiring dengan meningkatnya interaksi masyarakat dalam ruang maya terutama pada media sosial seperti Twitter dan Instagram yang menyediakan kanal ekspresi terbuka tanpa batasan ruang dan waktu. Tidak seperti bullying konvensional yang terbatas

secara fisik dan temporal, *cyberbullying* dapat terjadi kapan saja dan oleh siapa saja, bahkan secara anonim sehingga meningkatkan dampak psikologis yang ditimbulkan pada korban [40].

Tindakan *cyberbullying* dapat diklasifikasikan ke dalam beberapa bentuk ekspresi kekerasan verbal yang kompleks. Salah satu bentuk dominannya adalah ujaran bersifat abusive yaitu penggunaan bahasa kasar, menghina atau merendahkan martabat sesemanusia tanpa konteks yang sah [20]. Tindakan ini seringkali ditujukan untuk mempermalukan atau memprovokasi korban secara langsung di ruang publik digital. Selanjutnya, terdapat serangan terhadap individu yaitu narasi yang diarahkan secara personal kepada sesemanusia dengan maksud mencemarkan nama baik, membongkar informasi pribadi atau merendahkan karakter individu tersebut [21].

Selain menyerang secara personal, *cyberbullying* juga menyasar secara kolektif dalam bentuk serangan terhadap kelompok. Hal tersebut mencakup ujaran kebencian terhadap entitas sosial tertentu berdasarkan identitas kolektif seperti suku, bangsa, komunitas tertentu dan kelompok sosial lainnya [41]. Muncul juga bentuk *cyberbullying* berbasis identitas yang lebih spesifik seperti serangan terhadap agama, ras, fisik dan gender. Ujaran kebencian terhadap agama kerap kali memanfaatkan simbol-simbol kepercayaan sebagai sasaran pelecehan, sedangkan serangan berbasis ras atau etnis membawa dimensi diskriminatif yang kental [42]. *Body shaming* muncul dalam bentuk komentar yang melecehkan penampilan fisik, sedangkan kekerasan berbasis gender sering menyasar perempuan dan kelompok rentan lainnya dalam bentuk stereotip, pelecehan seksual verbal atau delegitimasi peran sosial.

Dalam klasifikasi intensitasnya, tindakan *cyberbullying* tidak selalu memiliki tingkat dampak yang seragam. *Cyberbullying* dapat dikategorikan ke dalam tiga tingkatan yaitu lemah (*weak*), sedang (*moderate*) dan kuat (*strong*) [43]. Kategori lemah mencakup ungkapan yang bersifat pasif-agresif atau sarkastik ringan namun tetap merendahkan. Kategori *moderate* mencerminkan intensitas yang lebih eksplisit dengan penggunaan diksi menghina atau melecehkan secara

langsung, namun belum mengandung ancaman serius. Sementara itu, kategori kuat merujuk pada konten yang sangat ofensif termasuk ajakan bunuh diri, kekerasan ekstrem atau ancaman serius terhadap keselamatan seseorang.

Pemahaman klasifikasi jenis dan tingkat *cyberbullying* ini sangat penting dalam membangun sistem deteksi otomatis berbasis kecerdasan buatan. Dengan mengenali secara detail corak ujaran yang digunakan, sistem tidak hanya mampu mengidentifikasi keberadaan kekerasan verbal secara umum, tetapi juga dapat mengklasifikasikan jenis kekerasan serta tingkat bahayanya. Dalam konteks ini, pemodelan berbasis *deep learning* seperti BiLSTM menjadi sangat relevan untuk mengakomodasi kompleksitas dan konteks semantik dari berbagai bentuk ujaran yang digunakan dalam praktik cyberbullying.

2.2.4 CRISP-DM

CRISP-DM (*Cross-Industry Standard Process for Data Mining*) merupakan kerangka kerja standar yang digunakan dalam proses data mining. Framework ini membantu menyederhanakan struktur data yang kompleks untuk menghasilkan informasi dan wawasan yang berguna [44]. CRISP-DM terdiri dari enam tahapan utama yang membentuk sebuah siklus iteratif, yang memungkinkan penyesuaian dan perbaikan berkelanjutan sepanjang proses analisis data [45]

Gambar 2. 1 Diagram CRISP-DM [46]

1. Business Understanding

Tahap awal ini bertujuan untuk memahami secara menyeluruh tujuan bisnis atau penelitian. Identifikasi permasalahan utama dan sasaran akhir dilakukan di sini, yang kemudian menjadi dasar dalam menentukan pendekatan data mining yang sesuai dan menyusun perencanaan proyek [47].

2. Data Understanding

Pada tahap ini, data dikumpulkan dari berbagai sumber, diperiksa kualitas dan strukturnya, serta dianalisis untuk memahami karakteristiknya. Proses ini membantu memastikan bahwa data yang akan digunakan cukup representatif dan relevan terhadap tujuan analisis [45].

3. Data Preparation

Data yang telah dipahami kemudian dipersiapkan untuk digunakan dalam pemodelan. Tahapan ini meliputi pemilihan data yang relevan, pembersihan data dari nilai yang tidak lengkap atau tidak konsisten, serta pembuatan atribut baru jika diperlukan untuk mendukung analisis [45].

4. Modeling

Di tahap ini, teknik pemodelan yang sesuai dipilih dan diaplikasikan pada data yang telah dipersiapkan. Model dibangun dan dikalibrasi menggunakan parameter tertentu, dengan mempertimbangkan tujuan bisnis dan karakteristik data [45]

5. Evaluation

Model yang telah dibangun dievaluasi untuk menilai sejauh mana hasilnya memenuhi tujuan bisnis yang telah ditetapkan di awal. Jika hasilnya belum memadai, maka perlu dilakukan peninjauan ulang terhadap tahapan sebelumnya untuk dilakukan perbaikan [45]

6. Deployment

Tahap akhir berupa implementasi model ke dalam lingkungan nyata, seperti sistem aplikasi bisnis atau laporan hasil analisis. Pemantauan dan

perbaikan terhadap model dilakukan secara berkala untuk menyesuaikan dengan perubahan kebutuhan atau kondisi bisnis yang terjadi [45].

2.3 Framework dan Algoritma

2.3.1 Natural Language Processing (NLP)

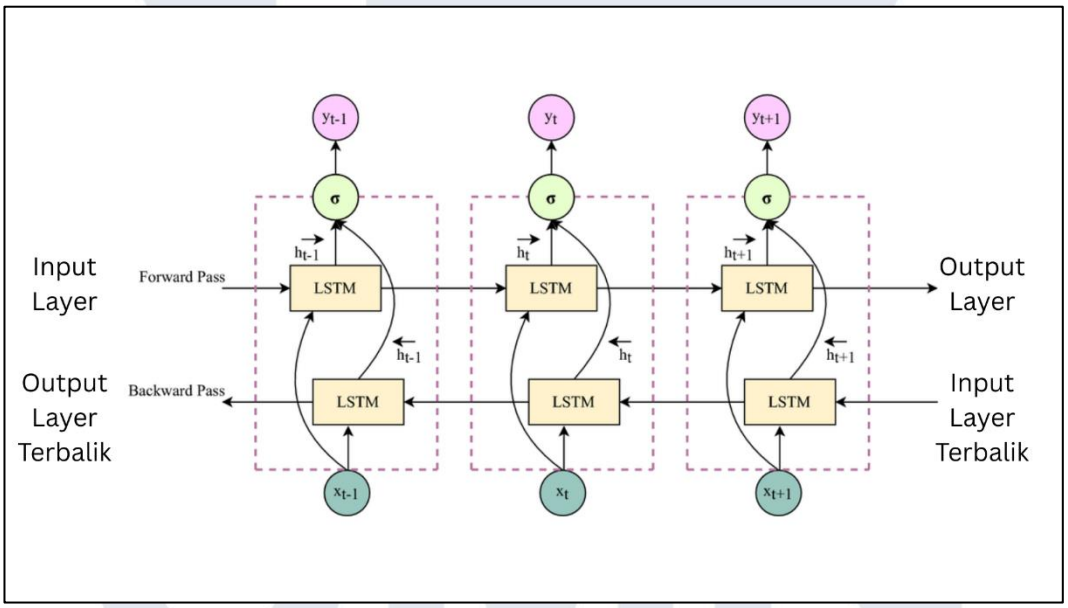
Natural Language Processing (NLP) adalah teknologi yang memungkinkan komputer untuk secara mendalam memahami, menganalisis dan menghasilkan teks atau suara yang digunakan oleh manusia dalam komunikasi mereka. NLP memanfaatkan berbagai teknik yang mencakup linguistik komputasi, statistik, *machine learning* dan *deep learning* untuk mengolah dan mengklasifikasikan data bahasa manusia [48]. Sebagai suatu cabang ilmu yang mengkombinasikan kecerdasan buatan dan pemahaman bahasa alami, NLP telah menjadi salah satu bidang yang paling menarik dalam dunia teknologi saat ini [49].

Salah satu komponen penting dalam penerapan analisis sentimen adalah NLP berbasis *machine learning* yang merupakan cabang dari kecerdasan buatan. Tujuan utama *machine learning* itu sendiri adalah memungkinkan sistem komputer belajar dari data dan menghasilkan prediksi atau keputusan tanpa harus diprogram secara eksplisit. Proses pembelajaran ini dilakukan dengan mengenali pola dan hubungan antar data dalam skala besar. Berdasarkan pendekatannya, *machine learning* dibagi menjadi empat kategori utama: supervised learning, *unsupervised learning*, *self-supervised learning* dan *reinforcement learning* [50]. Setiap pendekatan memiliki keunikan dalam hal pemrosesan dan penggunaan label data.

Dengan memproses bahasa alami secara mendalam, NLP dapat menyerap informasi penting dari teks tersebut, baik berupa opini, emosi, maupun konteks yang lebih luas. Kemampuan ini memungkinkan sistem untuk mengenali sentimen, mengidentifikasi topik utama [32] serta mengekstrak wawasan yang berguna dari berbagai jenis dokumen atau percakapan digital. Dengan demikian, NLP tidak hanya berfungsi sebagai alat pengolah bahasa, tetapi juga sebagai jembatan pemahaman antara manusia dan mesin yang mendukung pengambilan keputusan berbasis data tekstual.

2.3.2 Bidirectional Long Short Term Memory (BiLSTM)

Bidirectional Long Short-Term Memory (BiLSTM) merupakan pengembangan dari arsitektur Long Short-Term Memory (LSTM) yaitu salah satu jenis *Recurrent Neural Network (RNN)* yang dirancang untuk mengatasi permasalahan *vanishing gradient* pada pemrosesan data sekuensial . BiLSTM bekerja dengan cara membaca urutan data dalam dua arah, yakni dari depan ke belakang (*forward pass*) dan dari belakang ke depan (*backward pass*). Pendekatan ini memungkinkan model untuk memahami konteks penuh dari suatu kata dalam sebuah kalimat, baik dari sisi sebelumnya maupun sesudahnya. Dalam analisis teks seperti deteksi *cyberbullying*, kemampuan ini menjadi sangat penting karena makna dari suatu kata sering kali bergantung pada konteks yang muncul setelahnya, bukan hanya sebelumnya.



Gambar 2. 2 Struktur Algoritma Model Bi-LSTM [51]

Gambar 2.1 Menampilkan struktur algoritma BiLSTM yang terdiri dari dua jalur pemrosesan independen yang berjalan sejajar. "Forward Pass" (jalur atas) memproses urutan inpur dari kiri ke kanan (mulai dari xt−1, xt, hingga xt+1), dengan setiap blok LSTM menghasilkan *hidden state* (ht−1, ht, ht+1) yang diteruskan ke unit berikutnya dan ke lapisan output σ [51]. Sementara itu, "Backward Pass" (jalur bawah) memproses dalam urutan terbalik, dari kanan ke kiri (mulai dari xt+1, xt, hingga xt−1), dengan setiap blok LSTM juga menghasilkan

hidden state yang mengalir mundur. Output dari kedua arah ini (yaitu *hidden state* dari *forward pass* dan *backward pass*) kemudian akan digabungkan pada setiap langkah waktu, dan diteruskan ke lapisan selanjutnya, seperti simbol σ yang menghasilkan yt−1, yt, dan yt+1 sebagai keluaran akhir dari BiLSTM ini. Rumus matematis dasar dari unit LSTM yang digunakan dalam BiLSTM adalah sebagai berikut:

$$\mathrm{ft} = \sigma(\mathrm{Wf} \cdot [\mathrm{ht} - 1, \mathrm{xt}] + \mathrm{bf}) \qquad \text{(Forget gate)}$$

$$\mathrm{it} = \sigma(\mathrm{Wi} \cdot [\mathrm{ht} - 1, \mathrm{xt}] + \mathrm{bi}) \qquad \text{(Input gate)}$$

$$\mathrm{C{\sim}t} = \tanh(\mathrm{WC} \cdot [\mathrm{ht} - 1, \mathrm{xt}] + \mathrm{bC}) \qquad \text{(Candidate memory)}$$

$$\mathrm{Ct} = \mathrm{ft} * \mathrm{Ct} - 1 + \mathrm{it} * \mathrm{C{\sim}t} \qquad \text{(Cell state update)}$$

$$\mathrm{ot} = \sigma(\mathrm{Wo} \cdot [\mathrm{ht} - 1, \mathrm{xt}] + \mathrm{bo}) \qquad \text{(Output gate)}$$

$$\mathrm{ht} = \mathrm{ot} * \tanh(\mathrm{Ct}) \qquad \text{(Hidden state output)}$$

Keterangan:

- xt adalah input pada waktu ke-ttt,
- ht−1 adalah hidden state sebelumnya,
- Ct adalah cell state saat ini,
- σ adalah fungsi aktivasi sigmoid,
- tanh adalah fungsi aktivasi tanh,
- W dan b adalah bobot dan bias dari setiap gate.

Dalam BiLSTM, proses ini dilakukan dua kali secara simultan yaitu ke depan dan ke belakang, lalu hasil ht dari kedua arah digabungkan:

$$h_t^{BiLSTM} = concat(h_t^{forward}, h_t^{backward})$$

LSTM pertama akan membaca urutan kata dari awal ke akhir (forward), mempelajari bagaimana konteks terbentuk dari kata sebelumnya. Sementara itu, LSTM kedua berjalan dari akhir ke awal (backward), mempelajari konteks yang

muncul setelah kata tersebut. Misalnya, untuk kata “mengecewakan”, LSTM forward akan mempertimbangkan kata “sangat” dan “ini”, sedangkan LSTM backward akan melihat “dan” dan “membosankan”. Hasil hidden state dari dua arah ini digabungkan menjadi satu representasi konteks untuk kata tersebut.

Setiap waktu t, nilai aktivasi dari gate-gate LSTM, seperti input gate it, forget gate ft dan output gate ot, dihitung berdasarkan vektor input xt dan hidden state sebelumnya ht−1, sesuai dengan rumus matematis yang telah dijelaskan sebelumnya. Nilai-nilai ini digunakan untuk memperbarui *cell state* Ct dan menghasilkan *hidden state* hth_tht yang merepresentasikan informasi kontekstual dari kata tersebut. Proses ini terjadi secara simultan di kedua arah (forward dan backward), lalu digabungkan:

$$h_t^{BiLSTM} = concat(h_t^{forward}, h_t^{backward})$$

Hasil dari proses ini yaitu vektor htBiLSTM, digunakan oleh layer selanjutnya (misalnya dense layer atau softmax layer) untuk menentukan polaritas sentimen, dalam hal ini, kalimat kemungkinan besar diklasifikasikan sebagai negatif karena mengandung dua kata kunci negatif berturut-turut dengan penekanan emosional.

Contoh ini menunjukkan bagaimana BiLSTM tidak hanya memahami makna kata berdasarkan urutan sebelumnya, tetapi juga secara simultan mempertimbangkan konteks kata sesudahnya yang sangat krusial dalam menganalisis sentimen atau mendeteksi ujaran kebencian secara akurat terutama dalam bahasa Indonesia yang memiliki struktur fleksibel dan konteks yang kaya..

2.3.3 Focal Loss

Focal Loss adalah fungsi kerugian (loss function) yang dikembangkan untuk mengatasi masalah ketidakseimbangan kelas (class imbalance) yang sering terjadi dalam tugas klasifikasi, terutama pada skenario multi-label classification seperti deteksi cyberbullying [52]. Fungsi ini merupakan modifikasi dari Binary Cross Entropy (BCE) yang standar, dirancang khusus untuk memberikan penekanan lebih besar pada contoh-contoh yang sulit diprediksi dengan benar, sekaligus

menurunkan kontribusi dari contoh-contoh yang mudah diprediksi dengan tepat [53].

Ide utama di balik Focal Loss adalah memberikan bobot dinamis pada setiap sampel dalam proses pelatihan. Sampel yang mudah diklasifikasikan (misalnya, sampel dari kelas mayoritas yang sudah sering dikenali model) akan memiliki bobot kerugian yang lebih kecil, sementara sampel yang sulit (misalnya, sampel dari kelas minoritas atau yang ambigu) akan memiliki bobot kerugian yang lebih besar [54]. Hal ini dicapai dengan memperkenalkan faktor modulasi (1 - pt)^gamma ke dalam fungsi BCE, di mana pt adalah probabilitas prediksi model untuk kelas target, dan gamma adalah parameter yang mengontrol tingkat penekanan [55]. Nilai gamma yang lebih tinggi akan meningkatkan fokus model pada sampel yang lebih sulit.

2.3.4 Tokenization

Tokenization adalah salah satu tahapan fundamental dalam *Natural Language Processing* (NLP) yang melibatkan proses pemecahan teks menjadi unit-unit yang lebih kecil, yang disebut "token" [56]. Token-token ini dapat berupa kata, *subword*, karakter, atau frasa, tergantung pada granularitas yang diinginkan dan jenis *tokenizer* yang digunakan tok. Tujuan utama dari tokenisasi adalah untuk mengubah data teks mentah yang tidak terstruktur menjadi format yang dapat diproses dan dipahami oleh model *machine learning* atau *deep learning [55]*. Tanpa tokenisasi, model akan kesulitan dalam menganalisis dan mengekstraksi makna dari teks karena teks akan diperlakukan sebagai satu kesatuan yang besar.

Dalam konteks model *transformer* seperti BERT (termasuk IndoBERT yang digunakan dalam penelitian ini), *tokenization* memiliki peran yang sangat penting. Model-model ini bekerja dengan representasi numerik dari kata, sehingga setiap kata atau *subword* perlu diubah menjadi ID token yang unik [57]. Proses *tokenization* untuk IndoBERT melibatkan beberapa langkah, dimulai dengan penambahan token khusus seperti [CLS] di awal kalimat untuk merepresentasikan keseluruhan urutan teks pada tugas klasifikasi, dan [SEP] di akhir untuk menandai pemisahan atau akhir kalimat [58]. Setiap token ini kemudian diubah menjadi

embedding awal, yang juga diperkaya dengan informasi posisi (*Positional Encoding*) dan diatur fokus perhatiannya menggunakan *Attention Mask*.

Pentingnya tokenisasi juga terletak pada kemampuannya untuk menyeragamkan panjang *input* teks. Dalam penelitian ini, proses tokenisasi dilakukan menggunakan *BertTokenizer* dari IndoBERT dengan penyesuaian parameter berupa *padding* dan *truncation* agar setiap *input* memiliki panjang seragam dengan maksimal 100 token. *Padding* menambahkan token khusus (biasanya nol) untuk mencapai panjang maksimum yang ditentukan, sementara *truncation* memotong teks yang melebihi panjang maksimum. Hal ini memastikan bahwa semua *input* memiliki dimensi yang sama, yang merupakan persyaratan krusial untuk pelatihan model *deep learning* secara efisien. Dengan tokenisasi yang tepat, model dapat menerima *input* yang konsisten dan terstruktur, memungkinkan proses pembelajaran yang lebih stabil dan akurat dalam mendeteksi *cyberbullying*.

2.3.5 Confusion Matrix

Confusion Matrix adalah sebuah instrumen krusial yang digunakan untuk menilai efektivitas sebuah algoritma klasifikasi dalam konteks analisis sentimen. Alat ini memungkinkan evaluasi secara komprehensif terhadap ketepatan prediksi yang dihasilkan dengan membandingkan hasil klasifikasi model dengan data referensi yang sesungguhnya. Dalam representasi tabel seperti yang diperlihatkan pada Tabel 2.1, *Confusion Matrix* menyajikan visualisasi yang sistematis dan ringkasan performa model klasifikasi secara menyeluruh. Melalui matriks ini, kita dapat mengidentifikasi dengan jelas perbedaan dan kesesuaian antara nilai aktual yang merupakan kebenaran dasar dari dataset dan nilai prediksi yang dihasilkan oleh algoritma. Dengan demikian, penggunaan Confusion Matrix memungkinkan analisis mendalam mengenai kemampuan model dalam mengklasifikasikan data secara tepat dan efisien serta mengungkap aspek-aspek kesalahan prediksi yang mungkin terjadi [59].

Tabel 2 1 Confusion Matrix

Confusion Matrix		*Actual Data*	
		Positive	*Negative*
Prediction Result	*Positive*	*True Positive*	*False Negative*
	Negative	*False Positive*	*True Negative*

Melalui tabel Confusion Matrix yang telah disajikan, evaluasi terhadap hasil klasifikasi dapat dilakukan dengan memperhatikan empat komponen utama yaitu *True Positive* (TP), *False Positive* (FP), *False Negative* (FN) dan *True Negative* (TN). *True Positive* menunjukkan jumlah data aktual yang berlabel positif dan berhasil diprediksi dengan benar sebagai positif oleh model. Sebaliknya, *False Positive* merupakan jumlah data aktual negatif yang salah diklasifikasikan sebagai positif. *False Negative* menggambarkan data aktual positif yang secara keliru diprediksi sebagai negatif, sedangkan *True Negative* adalah jumlah data aktual negatif yang berhasil diidentifikasi dengan tepat sebagai negatif oleh model. Dengan memanfaatkan keempat nilai ini, berbagai metrik evaluasi dapat dihitung untuk mengukur kinerja metode yang dikembangkan secara menyeluruh. Setelah *Confusion Matrix* terbentuk, performa algoritma klasifikasi dapat dianalisis melalui parameter-parameter penting seperti akurasi, presisi, *recall* dan skor F1 yang secara kolektif memberikan gambaran mengenai keandalan model dalam mengklasifikasikan data serta kemampuannya dalam menangani kesalahan prediksi.

1. *Accuracy*

   Nilai *Accuracy* mencerminkan tingkat keakuratan model yang telah dibentuk dan diukur sebagai perbandingan antara data yang berhasil diklasifikasikan dengan benar (TP+TN) terhadap seluruh jumlah data

(TP+TN+FP+FN) [60]. Berikut merupakan cara perhitungan akurasi dapat diperjelas dengan berikut:

$$Accuracy = \frac{TP + TN}{Jumlah\ Data}$$

Rumus 2. 1 Accurracy.

2. *Precision*

Precision merupakan ukuran yang mengevaluasi sejauh mana model mampu mengenali kelas yang diminta dengan benar dalam perbandingan dengan semua hasil prediksi dari kelas tersebut. Dalam konteks ini, precision mencerminkan kemampuan model untuk memprediksi data aktual positif dari semua hasil prediksi yang dinyatakan sebagai kelas positif [60]. Rumus perhitungan precision merupakan sebagai berikut:

$$Precision = \frac{TP}{TP + FP}$$

Rumus 2. 2 Precision

3. *Recall*

Recall yang juga dikenal sebagai "sensitivitas" atau "true positive rate," mengukur sejauh mana model dapat mengenali kelas yang diminta dari seluruh data aktual yang termasuk dalam kelas tersebut. Dalam konteks ini, recall mencerminkan kemampuan model dalam memprediksi data aktual positif dari total keseluruhan data aktual positif dalam dataset [60]. Perhitungan recall dapat dinyatakan dalam rumus berikut:

$$Recall = \frac{TP}{TP + FN}$$

Rumus 2. 3 Recall

4. *F1 score*

F1-score atau F-Measure, merupakan suatu metrik yang memberikan gambaran tentang keseimbangan antara precision dan recall, khususnya ketika terjadi ketidakseimbangan dalam kelas data [60]. Rumus untuk menghitung F1-score merupakan sebagai berikut:

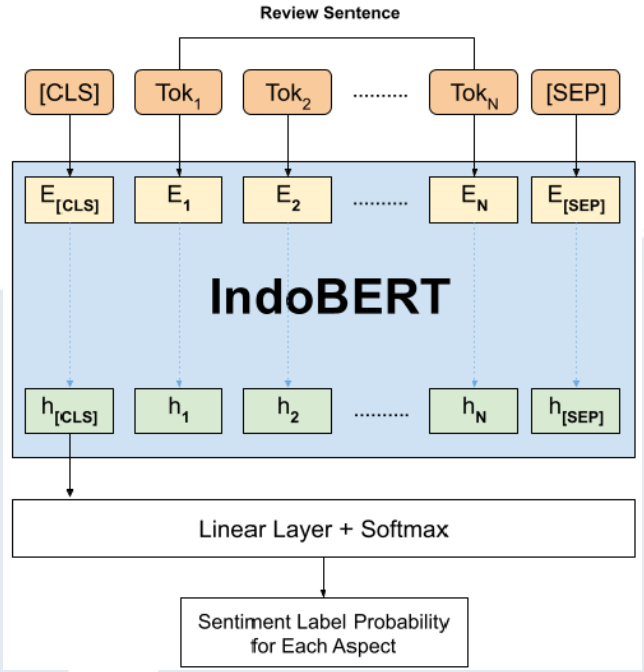
$$F1\ -score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Rumus 2. 4 F1 Score

2.3.6 IndoBERT

Dalam *Natural Language Processing*, representasi kata merupakan fondasi krusial untuk pemahaman konteks dan makna. Model *Bidirectional Encoder Representations from Transformers* (BERT) telah merevolusi bidang ini dengan kemampuannya menghasilkan *embedding* kontekstual yang kaya. Sebagai adaptasi dari arsitektur BERT, IndoBERT dirancang khusus untuk bahasa Indonesia, mengatasi tantangan leksikal dan sintaksis yang unik pada bahasa tersebut [61].

Proses di IndoBERT diawali dengan prapemrosesan input yang melibatkan tokenisasi dan penambahan token khusus seperti [CLS] (Classification Token) di awal kalimat untuk representasi keseluruhan sequence pada tugas klasifikasi, serta [SEP] (Separator) di akhir untuk menandai pemisahan atau akhir kalimat. Setiap token ini kemudian diubah menjadi embedding awal, yang juga diperkaya dengan informasi posisi (Positional Encoding) dan diatur fokus perhatiannya menggunakan Attention Mask. Embedding ini selanjutnya menjadi masukan bagi model IndoBERT. Pelatihan IndoBERT yang dilakukan pada korpus teks bahasa Indonesia yang besar, memungkinkan model untuk menangkap nuansa bahasa, idiom, dan pola linguistik yang spesifik, menghasilkan hidden state kontekstual untuk setiap token, dengan h[CLS] secara khusus menangkap informasi gabungan dari seluruh kalimat untuk tugas klasifikasi.



Gambar 2. 3 Struktur Algoritma Model IndoBERT [61]

Secara konseptual, arsitektur IndoBERT dapat digambarkan seperti pada Gambar 2.1. Proses diawali dengan tahap prapemrosesan *Review Sentence* (kalimat ulasan) yang akan dianalisis. Kalimat ini kemudian di tokenisasi dan diperkaya dengan token khusus: [CLS] di awal kalimat yang merepresentasikan seluruh sequence untuk tugas klasifikasi dan [SEP] di akhir yang menandai pemisahan kalimat. Setiap token (termasuk [CLS] dan [SEP]) akan diubah menjadi embedding awal yang disebut token embedding, dilambangkan sebagai E[CLS], E_1, E_2, …, E_n, E[SEP].*Embedding* ini selanjutnya menjadi masukan bagi model IndoBERT. Di dalam model IndoBERT, terjadi proses transformasi kompleks melalui *multi-head self-attention* dan *feed-forward neural networks* yang menghasilkan *hidden state* kontekstual. Setiap *hidden state* hi (di mana i dari [CLS], 1,…,N, hingga [SEP]) mencerminkan representasi kontekstual dari token yang bersangkutan, dengan *hidden state* h[CLS] secara khusus menangkap informasi gabungan dari seluruh kalimat. Secara matematis, transformasi ini dapat direpresentasikan sebagai:

$$hi = IndoBERTEncoder(Ei, PositionalEncodingi, AttentionMaski)$$

Di mana Ei adalah *embedding* masukan untuk token ke-i, PositionalEncodingi menambahkan informasi posisi token dalam urutan kalimat,

dan AttentionMaski digunakan untuk mengatur fokus perhatian model. Setelah melalui serangkaian lapisan *encoder* IndoBERT, *hidden state* h[CLS] yang merupakan representasi agregat dari seluruh masukan, kemudian diteruskan ke lapisan *Linear Layer* yang diikuti dengan fungsi aktivasi *Softmax*. Softmax adalah fungsi aktivasi yang digunakan pada tugas klasifikasi multi-class (satu label dari beberapa kelas), yang mengubah output menjadi probabilitas terdistribusi secara normal (totalnya 1). Lapisan ini bertugas untuk memproyeksikan representasi *hidden state* ke dalam ruang probabilitas kelas target. Hasil akhir dari proses ini adalah *Sentiment Label Probability for Each Aspect* (probabilitas label sentimen untuk setiap aspek), yang dalam konteks penelitian ini akan diinterpretasikan sebagai probabilitas *cyberbullying* atau *non-cyberbullying*.

Penggunaan IndoBERT dalam penelitian deteksi teks cyberbullying ini bertujuan untuk menghasilkan embedding kata yang powerful, yang kemudian dapat menjadi masukan bagi lapisan Bidirectional Long Short-Term Memory (BiLSTM) untuk klasifikasi yang lebih akurat. Potensi IndoBERT sebagai model pre-trained untuk bahasa Indonesia telah banyak dieksplorasi dalam berbagai tugas NLP. Misalnya, penelitian sebelumnya menunjukkan efektivitas IndoBERT dalam analisis sentimen pada ulasan aplikasi layanan kesehatan, mencapai akurasi sebesar 96% [62]. Selain itu, IndoBERT juga terbukti unggul dalam tugas Aspect-Based Sentiment Analysis (ABSA) pada ulasan pelanggan berbahasa Indonesia, bahkan mengungguli model BERT multibahasa dan Word2Vec dalam representasi teks kontekstual [61]. Studi lain juga memanfaatkan IndoBERT untuk klasifikasi deteksi hoax dalam bahasa Indonesia, menunjukkan peningkatan kinerja dibandingkan BERT asli dan model multibahasa [63]. Konsistensi kinerja IndoBERT yang superior dalam berbagai tugas klasifikasi teks berbahasa Indonesia menjadi dasar pertimbangan penggunaannya dalam penelitian ini, mengingat kemampuannya dalam memahami konteks semantik dan hubungan antar kata dengan lebih baik.

2.4 Tools dan Software Penelitian

2.4.1 Google Colaboratory / Google Colab

Google Colaboratory atau yang biasa dikenal dengan Google Colab adalah platform pemrograman berbasis cloud yang dikembangkan oleh

Google, memungkinkan pengguna menjalankan kode Python langsung melalui browser tanpa perlu instalasi lokal [64]. Platform ini berbasis Jupyter Notebook dan menyediakan akses ke sumber daya komputasi seperti CPU, GPU, bahkan TPU secara gratis [65]. Manfaat utama dari Google Colab adalah kemudahannya dalam melakukan eksperimen data science, pelatihan model machine learning, serta kolaborasi daring antar pengguna dalam satu dokumen notebook yang sama [66]. Kelebihan lain yang ditawarkan Colab meliputi integrasi langsung dengan Google Drive, dukungan pustaka populer seperti NumPy, pandas, TensorFlow, dan visualisasi interaktif tanpa harus memasang lingkungan pengembangan tambahan [67].

2.4.2 Python

Python adalah bahasa pemrograman tingkat tinggi yang bersifat open-source, fleksibel, dan mudah dipahami, pertama kali dikembangkan oleh Guido van Rossum dan dirilis pada tahun 1991. Bahasa ini mendukung berbagai paradigma pemrograman seperti prosedural, berorientasi objek, dan fungsional, sehingga dapat digunakan untuk berbagai kebutuhan pengembangan perangkat lunak [68]. Salah satu keunggulan utama Python adalah sintaksnya yang sederhana dan bersifat human-readable, menjadikannya bahasa yang sangat ramah bagi pemula maupun profesional [69]. Manfaat Python sangat luas, mulai dari pengembangan web, pemrosesan data, hingga kecerdasan buatan dan pembelajaran mesin. Hal ini diperkuat oleh ekosistem pustaka yang sangat kaya seperti NumPy, Pandas, Matplotlib hingga TensorFlow dan Scikit-learn, yang memungkinkan pengguna menyelesaikan tugas komputasi kompleks secara efisien [70]. Python juga bersifat multiplatform, dapat dijalankan di berbagai sistem operasi tanpa banyak modifikasi kode, serta mudah diintegrasikan dengan bahasa lain seperti C/C++ atau Java [71].