BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Penelitian terkait analisis sentimen dengan pendekatan *Natural Language Processing (NLP)* terus berkembang seiring meningkatnya ketersediaan data dari media sosial, khususnya X. Beragam metode telah diterapkan untuk meningkatkan kualitas klasifikasi sentimen, termasuk model *deep learning* dan *hybrid deep learning*. *LSTM, BiLSTM*, serta kombinasi *CNN* dengan *LSTM* maupun *BiLSTM* merupakan beberapa metode yang sering dipakai karena berkemampuan membuahkan performa baik dalam memahami konteks teks dan mengelompokkan sentimen secara akurat.

Tabel 2.1 menampilkan beberapa penelitian terdahulu yang berhubungan dengan implementasi teknik analisis sentimen, terutama penelitian yang memakai model *Hybrid CNN-LSTM* dan *Hybrid CNN-BiLSTM*. Penelitian tersebut dipakai sebagai landasan teori sekaligus referensi dalam melihat perbandingan performa model pada analisis sentimen terkait isu perlindungan data pribadi. Berikut beberapa penelitian terdahulu yang dijadikan kajian dalam penelitian ini:

Tabel 2. 1 Penelitian Terdahulu

No	Judul dan Peneliti	Nama Jurnal	Metode	Hasil Penelitian
1	*Sentiment Analysis of Covid-19 Twitter Data using Deep Learning Algorithm* [8]	*Procedia Computer Science* (2024)	*ANN, LSTM*	Model *LSTM* mencapai akurasi sebanyak 96% dan membuktikan bahwa sebagian besar *tweet* terkait COVID-19 memiliki sentimen

No	Judul dan Peneliti	Nama Jurnal	Metode	Hasil Penelitian
				positif atau netral, sedangkan metode *ANN* hanya mencapai akurasi sebanyak 76%.
2	*Sentiment Analysis using Bidirectional LSTM Network* [9]	*Procedia Computer Science* (2023)	*RNN*, *Bi-LSTM*	Model *Bidirectional LSTM (Bi-LSTM)* membuktikan performa yang sangat baik dalam analisis sentimen dengan akurasi mencapai 91,4%, serta memiliki potensi untuk dikembangkan dalam klasifikasi sentimen yang lebih detail, termasuk identifikasi emosi seperti senang, sedih, dan marah.
3	*Hybrid singular spectrum analysis and LSTM modeling for daily precipitation forecasting* [11]	*Geodesy and Geodynamics* (2026)	*SSA-LSTM*	Model *hybrid SSA-LSTM* mampu meningkatkan akurasi prediksi secara signifikan dibandingkan model tunggal, dengan nilai R^2

No	Judul dan Peneliti	Nama Jurnal	Metode	Hasil Penelitian
				mencapai 0,988–0,990 serta penurunan nilai *RMSE* dan *MAE* yang signifikan. Hal ini membuktikan bahwa kombinasi *SSA* dan *LSTM* efektif dalam mengekstraksi fitur temporal dan menangkap hubungan nonlinier pada data.
4	*Unraveling user perceptions and biases: A comparative study of ML and DL models for exploring twitter sentiments towards ChatGPT* [10]	*Journal of Engineering Research* (2025)	*Bi-LSTM, GRU, LSTM, CNN, Random Forest (RF), SVM, CNN-BiLSTM*	Model *hybrid CNN-BiLSTM* dengan *embedding FastText* menghasilkan akurasi tertinggi sebanyak 96,59% dibandingkan metode lainnya. Secara umum, model *deep learning* membuktikan performa lebih baik dibandingkan *machine learning* tradisional. Hasil

No	Judul dan Peneliti	Nama Jurnal	Metode	Hasil Penelitian
				analisis sentimen membuktikan mayoritas *tweet* bernada negatif (terkait privasi, bias, dan pekerjaan), diikuti sentimen positif (membantu dan canggih), serta sebagian netral.
5	*Twitter sentiment analysis using conditional generative adversarial network* [4]	*International Journal of Cognitive Computing in Engineering* (2024)	*CNN, LSTM, Bi-LSTM, CGAN*	Model *CGAN* mencapai akurasi 93,33%, lebih tinggi dibanding *CNN*, *LSTM*, dan *Bi-LSTM*. *CGAN* juga unggul dalam *precision*, *recall*, dan *F1-score*, serta lebih cepat dalam klasifikasi *tweet*.
6	*Social media-based COVID-19 sentiment classification model using Bi-LSTM* [5]	*Expert Systems with Applications* (2023)	*Bi-LSTM*	Model *Bi-LSTM* mampu mengelompokkan sentimen (positif, negatif, netral) dari data media sosial terkait COVID-19 dengan akurasi tinggi (86–97%). Hasil

No	Judul dan Peneliti	Nama Jurnal	Metode	Hasil Penelitian
				membuktikan *Bi-LSTM* lebih unggul dibanding metode lain seperti *CNN*, *LSTM*, dan *Naive Bayes*. Model ini efektif untuk memahami opini publik, mendeteksi misinformasi, dan mendukung pengambilan keputusan kebijakan kesehatan.
7	*Stratification of Depressed and Non-Depressed Texts from Social Media using LSTM and its Variants* [12]	*Procedia Computer Science* (2024)	*Classic LSTM, Stacked LSTM, Bi-LSTM, GRU, Bi-GRU*	Model *Classic LSTM* dan *GRU* mencapai akurasi tertinggi (95%). *Bi-LSTM* dan *Bi-GRU* unggul dalam presisi, tetapi tidak lebih baik dalam akurasi. *GRU* lebih efisien untuk dataset besar dan kalimat panjang.
8	*A hybrid deep learning approach for detecting sentiment polarities and knowledge graph*	*Decision Analytics Journal* (2023)	*Hybrid CNN-LSTM*	Model *hybrid CNN-LSTM* mencapai akurasi 94% dalam mendeteksi

No	Judul dan Peneliti	Nama Jurnal	Metode	Hasil Penelitian
	representation on monkeypox tweets [7]			sentimen (positif, negatif, netral). *Knowledge graph* berhasil memetakan hubungan antar topik (gejala, vaksin, respon pemerintah) dalam percakapan X tentang *monkeypox*
9	*Transforming sentiment analysis for e-commerce product reviews: Hybrid deep learning model with an innovative term weighting and feature selection* [13]	*Information Processing & Management* (2024)	*EGJO-LSTM*	Model *EGJO-LSTM* mampu meningkatkan performa analisis sentimen ulasan produk e-commerce dengan pengembangan presisi sekitar 25% dan akurasi sekitar 32% dibandingkan metode lain. Selain itu, model ini mampu mengelompokkan sentimen ke dalam lima kategori, yaitu sangat positif, positif, netral, negatif, dan sangat negatif.

No	Judul dan Peneliti	Nama Jurnal	Metode	Hasil Penelitian
10	*Sentiment analysis in multilingual context: Comparative analysis of machine learning and hybrid deep learning models* [14]	*Heliyon* (2023)	*Bi-LSTM, SVM*	Metode *SVM* menghasilkan akurasi tertinggi, yaitu sebanyak 82,56% untuk bahasa Inggris dan 86,43% untuk bahasa Bangla. Sementara itu, *Bi-LSTM* membuktikan performa terbaik di antara model *deep learning* dengan akurasi sebanyak 78,10% (Inggris) dan 83,72% (Bangla). Model *Conv1D* murni memiliki performa paling rendah dibandingkan metode lainnya.
11	*Application of bidirectional LSTM deep learning technique for sentiment analysis of COVID-19 tweets: post-COVID vaccination era* [6]	*Journal of Electrical Systems and Information Technology* (2023)	*Bi-LSTM*	Akurasi model 78,29%; Sentimen positif 49,93% dan negatif 50,06%; opini publik hampir seimbang, membuktikan adanya dukungan sekaligus keraguan

No	Judul dan Peneliti	Nama Jurnal	Metode	Hasil Penelitian
				atas vaksin COVID-19.

Berdasarkan Tabel 2.1, berbagai penelitian sebelumnya mengungkapkan bahwa analisis sentimen pada media sosial, terutama X, telah banyak dipakai untuk mengetahui pandangan masyarakat atas berbagai isu, seperti COVID-19, ChatGPT, *monkeypox*, kesehatan mental, hingga ulasan produk *e-commerce*. Fokus utama dari penelitian tersebut adalah meningkatkan hasil pemilahan sentimen pada data teks yang bersifat tidak terstruktur. Penelitian yang memakai *Artificial Neural Network (ANN)* dan *Long Short-Term Memory (LSTM)* membuktikan bahwa metode *LSTM* mendapatkan tingkat akurasi lebih tinggi, yaitu sebanyak 96%, sedangkan dengan *ANN* hanya mencapai 76% [8]. Berikutnya, pengembangan metode *Bidirectional Long Short-Term Memory (BiLSTM)* juga dinilai lebih efektif dibandingkan dengan *Recurrent Neural Network (RNN)* dengan akurasi mencapai 91,4% [9]. Penelitian lain memakai model *hybrid SSA-LSTM* memperlihatkan adanya pengembangan akurasi prediksi yang cukup signifikan dengan nilai R^2 sebanyak 0,988–0,990 [11]. Selain itu, model *Hybrid CNN-BiLSTM* mendapatkan tingkat akurasi tertinggi sebanyak 96,59% dibandingkan dengan beberapa metode lain seperti *CNN*, *LSTM, BiLSTM, GRU, Random Forest,* dan *SVM* [10]. Model *CGAN* juga memberikan kinerja yang lebih unggul dibandingkan dengan *CNN, LSTM*, dan *BiLSTM* dengan tingkat kinerja sebanyak 93,33% [4]. Pada penelitian terkait COVID-19, metode *BiLSTM* mampu menghasilkan akurasi antara 86–97% [5], sedangkan penelitian mengenai teks depresi membuktikan bahwa metode *Classic LSTM* dan *GRU* mendapatkan akurasi terbaik sebanyak 95% [12]. Penelitian memakai model *Hybrid CNN-LSTM* pada *monkeypox tweets* mendapatkan tingkat akurasi sebanyak 94% [7]. Sementara itu, model *EGJO-LSTM* pada analisis ulasan *e-commerce* mampu meningkatkan akurasi hingga sekitar 32% [13]. Penelitian terdahulu melaporkan bahwa *SVM* mampu mengungguli *BiLSTM* dalam *multilingual sentiment analysis* [14], sedangkan *BiLSTM* pada analisis sentimen terkait vaksin COVID-19 menghasilkan akurasi

sebanyak 78,29% [6]. Beralaskan hasil penelitian tersebut, bisa disimpulkan bahwa metode *deep learning* dan *hybrid deep learning* adalah pendekatan yang efektif untuk menganalisis sentimen, terutama pada data teks yang didapatkan dari media sosial.

Mengacu pada berbagai penelitian sebelumnya, penelitian ini memanfaatkan pendekatan *Natural Language Processing (NLP)* untuk melakukan analisis sentimen memakai data yang berasal dari X. Penelitian berfokus pada pengelompokan sentimen masyarakat terkait isu perlindungan data pribadi di Indonesia. Model yang dipakai adalah *Hybrid CNN-LSTM* dan *Hybrid CNN-BiLSTM* karena kedua metode tersebut telah membuktikan kemampuan yang baik dalam tugas kategorisasi sentimen pada penelitian terdahulu [7], [10]. Penelitian ini melibatkan beberapa tahapan pengolahan data awal, yaitu *cleaning, case folding*, dan *tokenization* sebelum metode pemodelan dijalankan [8]. Setelah data selesai dimetode, data tersebut dipilah lagi sehingga mendapatkan dua bagian, yaitu data latih dan data uji. Untuk mengetahui seberapa baik model bekerja, dipakai metrik *accuracy*, *precision*, *recall*, dan *F1-score* sebagai acuan dalam membandingkan hasil klasifikasi yang didapat.

Adapun perbedaan dari penelitian ini terletak pada fokus analisis sentimen atas isu perlindungan data pribadi di Indonesia yang diambil dari data X dan relevan dengan konteks lokal. Berbeda dengan penelitian sebelumnya yang sebagian besar membahas isu kesehatan seperti COVID-19 [8], [5], [6], isu penyakit seperti *monkeypox* [7], maupun isu teknologi seperti ChatGPT [10]. Penelitian ini menitikberatkan pada persepsi masyarakat atas keamanan data pribadi. Selain itu, penelitian ini secara khusus membandingkan performa model *Hybrid CNN-LSTM* dan *Hybrid CNN-BiLSTM* untuk mengetahui model yang paling optimal dalam mengelompokkan sentimen positif, negatif, dan netral. Perbandingan langsung antara kedua model *hybrid* tersebut pada isu perlindungan data pribadi masih belum banyak dijalankan, sehingga penelitian ini dimaksudkan dapat menyampaikan peran baru pada bidang analisis sentimen berbasis media sosial.

2.2 Teori yang berkaitan

2.2.1 Text Mining

Text mining ialah salah satu cara yang dipakai untuk mengolah dan menganalisis data teks supaya bisa mendapatkan informasi yang berguna dari data yang tidak memiliki struktur yang jelas. Pada media sosial seperti X, teks yang ada biasanya memakai berbagai bentuk bahasa, singkatan, tanda baca, dan kata-kata yang tidak formal, sehingga membutuhkan metode khusus dalam pemrosesan. Dalam penelitian analisis sentimen, *text mining* dipakai untuk membarui data teks menjadi bentuk yang lebih rapi, sehingga analisis bisa dijalankan dengan lebih baik [14].

Tahapan dalam *text mining* umumnya mencakup pengumpulan data, *preprocessing*, representasi data dalam bentuk numerik, klasifikasi, dan evaluasi hasil analisis. Pada tahap *preprocessing*, dijalankan serangkaian metode seperti *cleaning*, *case folding*, tokenisasi, serta representasi teks ke dalam bentuk numerik sebelum metode pemodelan. Penelitian terdahulu membuktikan bahwa kualitas *preprocessing* berpengaruh atas kinerja model karena dapat meminimalkan data yang tidak relevan dan meningkatkan kualitas fitur yang dipakai [13].

Selain itu, *text mining* juga sering diterapkan dalam analisis sentimen karena dapat membantu metode pengenalan opini masyarakat secara otomatis dari data teks dalam jumlah besar. Analisis sentimen ialah bagian dari *text mining* yang dipakai untuk mengenali pendapat pemakai, apakah positif, negatif, atau netral. Penerapan *text mining* pada data X dinilai efektif untuk mengetahui pola sentimen masyarakat sekaligus memetakan topik pembahasan atas suatu isu tertentu [7].

Pemakaian *text mining* dalam penelitian ini bertujuan untuk mengolah data opini masyarakat terkait keamanan data pribadi di Indonesia yang bersumber dari X. Dengan melalui beberapa langkah pengolahan, data teks yang tidak memiliki struktur bisa diganti menjadi data yang lebih rapi dan siap dipakai untuk menganalisis sentimen memakai model *Hybrid CNN-LSTM* dan

Hybrid CNN-BiLSTM untuk mengetahui kecenderungan opini publik atas isu perlindungan data pribadi.

2.2.2 Natural Language Processing (NLP)

Natural Language Processing (NLP) merupakan area dalam *Artificial Intelligence* yang mempelajari teknik agar komputer bisa mengerti dan memproses bahasa manusia secara otomatis. Dengan *NLP*, sistem dapat mengenali pola linguistik, memahami makna suatu kata, serta menganalisis hubungan antar kata dalam sebuah kalimat. Dalam mengolah teks, *NLP* dipakai untuk mengubah bahasa alami menjadi bentuk yang bisa dibaca dan dipahami oleh algoritma *machine learning* serta *deep learning*. Oleh karena itu, *NLP* sering dipakai dalam analisis sentimen untuk membantu sistem mengenali dan mengelompokkan pendapat ada dalam data teks secara otomatis [15].

Dalam analisis sentimen, *NLP* memiliki peran penting dalam mengolah data teks agar dapat dimetode oleh model komputasi. Beberapa teknik NLP yang umum dipakai mencakup *cleaning*, *case folding*, tokenisasi, dan normalisasi teks. Namun, pada pengkajian ini metode preprocessing yang diterapkan meliputi *cleaning, case folding*, serta tokenisasi sesuai dengan kebutuhan dataset. Tahapan tersebut dijalankan untuk membersihkan data teks, mengurangi noise, serta mengembangkan mutu representasi data sehingga hasil klasifikasi sentimen dapat menjadi lebih optimum. Beberapa penelitian juga membuktikan bahwa *preprocessing* berbasis *NLP* sangat berpengaruh atas performa model klasifikasi sentimen karena mampu menghasilkan fitur yang lebih relevan dan efektif [16].

Selain dipakai untuk *preprocessing*, *NLP* juga berfungsi untuk memahami konteks kalimat secara lebih mendalam, terutama pada data media sosial seperti X yang memiliki gaya bahasa tidak baku, banyak singkatan, serta pemakaian simbol yang beragam. Dengan adanya *NLP*, sistem dapat mengenali pola sentimen pemakai secara otomatis dan membagi anggapan ke dalam golongan positif, negatif, atau netral. Sebab itu, *NLP* menjadi salah satu metode

yang sering dipakai dalam penelitian analisis sentimen berbasis media sosial [17].

2.2.3 Analisis Sentimen

Sebagai salah satu penerapan *NLP*, analisis sentimen berusaha untuk mengenali dan mengelompokkan pendapat yang terdapat dalam data teks. Metode ini dijalankan dengan menentukan polaritas sentimen suatu teks ke dalam kategori positif, negatif, atau netral. Karena bisa mengolah data tidak teratur secara otomatis, analisis sentimen sering dipakai untuk memahami pendapat dan respons masyarakat atas suatu isu tertentu [17].

Analisis sentimen sering dipakai untuk mengetahui pendapat masyarakat di berbagai bidang, seperti bisnis, politik, kesehatan, teknologi, dan pemerintahan. Di media sosial X, metode ini dipakai untuk mengetahui bagaimana sentimen pemakai atas topik tertentu yang sedang tren. Karakteristik data media sosial yang bersifat *real-time* membuat analisis sentimen menjadi pendekatan yang efektif dalam memantau opini dan respons masyarakat atas berbagai isu [15].

Analisis sentimen tidak hanya membantu mengetahui pendapat masyarakat, tetapi juga menjadi bantuan dalam mengambil keputusan. Informasi yang diperoleh dari metode klasifikasi sentimen dapat dimanfaatkan sebagai dasar evaluasi berbagai kebijakan, layanan, maupun isu tertentu [18]. Pada bidang pemerintahan, analisis sentimen membantu memahami tanggapan publik atas kebijakan yang diterapkan, sedangkan dalam dunia bisnis metode ini dipakai untuk mengetahui pendapat dan tingkat kepuasan pelanggan atas produk atau layanan yang ditawarkan.

Dalam analisis sentimen, teks biasanya dibagi menjadi tiga jenis, yaitu sentimen positif, negatif, dan netral. Metode pengelompokan tersebut dijalankan dalam beberapa langkah, yaitu mulai dari *preprocessing, feature extraction*, hingga pemakaian algoritma *machine learning* maupun *deep learning* untuk metode pengelompokan. Beberapa penelitian juga menjelaskan

bahwa model *hybrid deep learning* mampu meningkatkan hasil analisis sentimen karena dapat menggabungkan kemampuan ekstraksi fitur dan interpretasi konteks kalimat dengan lebih baik [13]. Oleh karena itu, model seperti *Hybrid CNN-LSTM* dan *Hybrid CNN-BiLSTM* sering dipakai dalam analisis sentimen karena mampu memberikan hasil pengelompokan yang lebih baik pada data teks media sosial [10], [7].

Pada penelitian ini, analisis sentimen dipakai untuk memahami pendapat masyarakat mengenai isu keamanan data pribadi di Indonesia berdasarkan data yang berasal dari X. Dengan memakai teknik analisis sentimen, data teks yang belum teratur dapat diganti menjadi informasi yang lebih teratur, sehingga lebih mudah dianalisis memakai model *Hybrid CNN-LSTM* dan *Hybrid CNN-BiLSTM* untuk melihat kecenderungan sentimen masyarakat atas isu perlindungan data pribadi.

2.2.4 Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) ialah salah satu teknik *deep learning* yang pada awalnya sering dipakai untuk pengolahan citra. Namun, seiring perkembangan teknologi, *CNN* juga mulai dipakai dalam pengolahan teks, termasuk analisis sentimen. *CNN* bekerja dengan melakukan ekstraksi fitur secara otomatis melalui metode konvolusi, sehingga mampu mengidentifikasi pola kata, frasa, serta hubungan antar kata dalam teks. Kemampuan itu membuat *CNN* sangat efektif dalam mengolah data teks yang tidak teratur dan mempunyai tingkat kesulitan yang tinggi [10].

Dalam analisis sentimen, *CNN* dipakai sebagai *feature extractor* yang membantu model mengenali informasi penting dari data teks, seperti pola lokal antar kata yang memiliki hubungan makna tertentu. Salah satu kelebihan *CNN* adalah kemampuannya menangkap local features secara lebih efisien serta membantu mengurangi kompleksitas data sebelum dimetode oleh model sekuensial seperti *LSTM* atau *BiLSTM*. Sebagian penelitian membuktikan bahwa penerapan *CNN* pada model *hybrid* dapat mengembangkan hasil

klasifikasi sentimen karena metode ekstraksi fitur menjadi lebih efektif dan optimal [7].

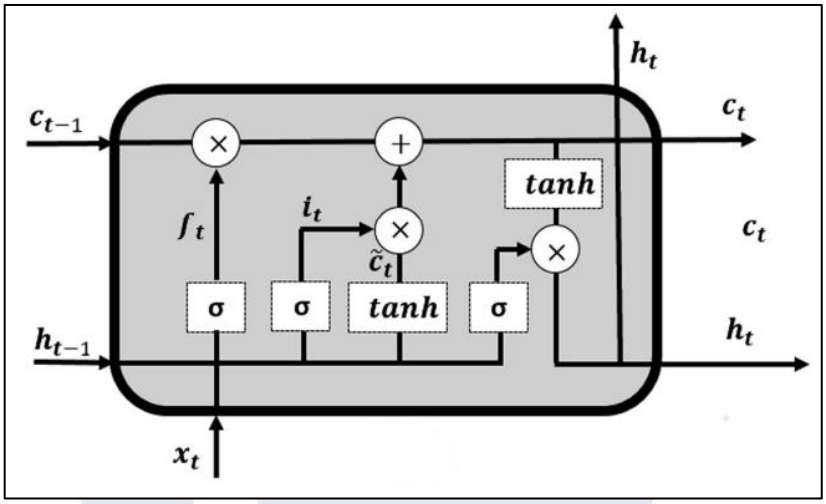
CNN sering dikombinasikan dengan LSTM dan BiLSTM untuk membangun model deep learning hybrid, seperti *Hybrid CNN-LSTM* dan *Hybrid CNN-BiLSTM*. Pada arsitektur tersebut, *CNN* bertugas melakukan ekstraksi fitur dari data teks, sementara *LSTM* atau *BiLSTM* dipakai untuk memahami konteks dan dependensi antar kata dalam suatu urutan kalimat. Beberapa penelitian membuktikan bahwa pendekatan *hybrid* ini dapat memberikan akurasi yang lebih tinggi dibandingkan model tunggal karena mengintegrasikan kemampuan kedua metode secara bersamaan [10].

Pada penelitian ini, *CNN* dipakai sebagai komponen utama dalam model *Hybrid CNN-LSTM* dan *Hybrid CNN-BiLSTM* pada metode analisis sentimen data X terkait isu perlindungan data pribadi di Indonesia. *CNN* berfungsi untuk melakukan ekstraksi fitur pada teks sehingga model dapat mengenali sentimen positif, negatif, maupun netral dengan tingkat akurasi yang lebih baik.

2.2.5 Long Short-Term Memory (LSTM)

LSTM merupakan salah satu bentuk arsitektur *RNN* yang dibuat untuk menangani kekurangan *RNN* konvensional, khususnya masalah *vanishing gradient* saat mengolah data berurutan. Model ini memanfaatkan *memory cell* dan tiga komponen utama, yaitu *input gate*, *forget gate*, dan *output gate*, untuk mengelola aliran informasi selama metode pembelajaran. Dengan cara itu, *LSTM* mampu menyimpan informasi penting secara jangka panjang dan meminimalkan dampak dari informasi yang tidak relevan [8]. Dalam bidang *NLP*, *LSTM* banyak dipakai sebab keahliannya dapat mengerti konteks berdasarkan rangkaian kata pada teks. Kemampuan tersebut membuat *LSTM* efektif untuk diterapkan pada analisis sentimen, klasifikasi teks, *machine translation*, dan *speech recognition*. Pada data X, *LSTM* mampu mengidentifikasi pola sentimen dengan baik sehingga menghasilkan performa klasifikasi yang kompetitif [12].

Hasil penelitian terdahulu membuktikan bahwa model *LSTM* mampu membagikan performa yang tinggi dalam tugas analisis sentimen. Salah satu penelitian pada data X terkait COVID-19 melaporkan bahwa model ini mendapatkan akurasi hingga 96% dalam metode klasifikasi sentimen [8].



Gambar 2. 1 Arsitektur *Single Cell Long Short-Term Memory (LSTM)* [12]

Gambar 2.1 membuktikan bahwa *LSTM* memiliki tiga komponen utama, yaitu *forget gate*, *input gate*, dan *output gate*, yang berperan dalam mengontrol metode penyimpanan, pembaruan, dan pengeluaran informasi pada jaringan. Dengan mekanisme tersebut, *LSTM* mampu mempelajari dependensi jangka panjang pada data sekuensial sehingga banyak dipakai dalam analisis sentimen. Adapun formulasi matematis yang dipakai pada setiap *gate* dijelaskan sebagai berikut [12]:

Forget gate dipakai untuk memastikan informasi yang ingin dijaga atau dihapus dari *cell state* sebelumnya:

$$f_t = sigmoid\ (W_f * [H_{t-1}, X_t] + b_f$$

Rumus 1 *Forget Gate* pada *LSTM*

Input gate dipakai untuk memastikan informasi baru yang akan disimpan:

$$i_t = sigmoid\ (W_f*[H_{t-1},X_t] + b_i)$$

Rumus 2 *Input Gate* pada *LSTM*

Output gate dipakai untuk memastikan informasi yang akan dikeluarkan:

$$c_t = f_t * C_{t-1} + i_t * \tanh(W_c * [\ H_{t-1},X_t] + b_c)$$

Rumus 3 *Output Gate* pada *LSTM*

Pembaruan *cell state* dirumuskan sebagai:

$$o_t = sigmoid\ (W_o * [\ W_{t-1}, X_t] + b_o)$$

Rumus 4 Pembaruan *Cell State* pada *LSTM*

Output akhir dari *LSTM (hidden state)* dirumuskan sebagai berikut:

$$H_t = o_t\ * \tanh(C_t)$$

Rumus 5 *Hidden State* pada *LSTM*

Keterangan dari simbol-simbol yang dipakai pada rumus LSTM adalah sebagai berikut:

- f_t : *forget gate*, yaitu nilai yang memastikan informasi yang dipertahankan atau dihapus dari *cell state* sebelumnya
- i_t : *input gate*, yaitu nilai yang memastikan informasi baru yang akan disimpan ke dalam *cell state*
- o_t : *output gate*, yaitu nilai yang menentukan informasi yang akan dikeluarkan sebagai output
- c_t : *cell state* pada waktu ke-t
- C_{t-1} : *cell state* pada waktu sebelumnya
- H_t : *hidden state* pada waktu ke-t (*output LSTM*)
- H_{t-1} : *hidden state* pada waktu sebelumnya

- X_t : input pada waktu ke-t
- W_f, W_c, W_t, W_o : bobot (*weight*) pada masing-masing *gate*
- b_f, b_i, b_c, b_o: bias pada masing-masing *gate*
- σ (*sigmoid*) : fungsi aktivasi yang membangun nilai antara 0 dan 1
- tanh : fungsi aktivasi yang membangun nilai antara -1 hingga 1
- [$H_{t-1,}X_t$] : metode penggabungan (*concatenation*) antara *hidden state* sebelumnya dan input saat ini

Selain dipakai sebagai model tunggal, *LSTM* juga sering dikombinasikan dengan *Convolutional Neural Network (CNN)* untuk membentuk model *Hybrid CNN-LSTM*. Dalam kombinasi ini, *CNN* bertugas untuk mengekstrak pola penting dari teks, sementara *LSTM* dipakai untuk memahami urutan kata dalam kalimat. Kombinasi tersebut membantu meningkatkan hasil analisis sentimen karena metode ekstraksi fitur dan pemahaman konteks dijalankan secara lebih optimal. Penelitian membuktikan bahwa model *Hybrid CNN-LSTM* berhasil mendapatkan tingkat akurasi sebanyak 94% dalam mengkaji sentimen data X [7].

Dalam penelitian ini, *LSTM* dipakai sebagai bagian utama dalam model *Hybrid CNN-LSTM* untuk mengelompokkan sentimen data X yang berhubungan dengan isu penjagaan data pribadi di Indonesia. Diharapkan dalam memakai LSTM, kemampuan model dalam memahami urutan kalimat secara kontekstual akan meningkat, sehingga hasil klasifikasi sentimen menjadi lebih tepat.

2.2.6 Bidirectional Long Short-Term Memory (BiLSTM)

Bidirectional Long Short-Term Memory (BiLSTM) adalah ekspansi dari *Long Short-Term Memory (LSTM)* yang dibuat agar model lebih mampu memahami konteks dari data berurutan. Terdapat perbedaan dengan *LSTM* yang mengolah data dari satu arah, *BiLSTM* mampu mengolah data secara dua arah sekaligus, yaitu *forward* dan *backward*. Pendekatan ini mengharuskan model mendapatkan informasi yang lebih lengkap karena makna suatu kata sering

terpengaruh oleh kata-kata yang muncul sebelum maupun sesudahnya. Dengan memanfaatkan dua arah pemrosesan tersebut, *BiLSTM* mampu menangkap hubungan antar kata secara lebih komprehensif dan menghasilkan representasi konteks yang lebih baik [9]. Dalam implementasinya, keluaran dari metode *forward* dan *backward* digabungkan untuk membentuk representasi informasi yang lebih optimal.

Pada area *Natural Language Processing (NLP)*, *BiLSTM* sering dipakai untuk berbagai pekerjaan dalam pengolahan teks, terutama analisis sentimen, karena kemampuannya yang baik dalam memahami hubungan antar kata berdasarkan. Karakteristik tersebut membuat *BiLSTM* efektif dipakai pada data sekuensial seperti *tweet*, di mana pemahaman atas keseluruhan konteks kalimat sangat diperlukan. Beberapa penelitian menyebutkan bahwa *BiLSTM* mampu menyampaikan hasil yang lebih unggul dibandingkan *LSTM* biasa dan beberapa metode *machine learning* tradisional pada kasus analisis sentimen tertentu [5].

BiLSTM tidak hanya dipakai sebagai model tunggal, tetapi juga sering dikombinasikan dengan *CNN* dalam bentuk *Hybrid CNN-BiLSTM*. Dalam arsitektur ini, *CNN* berfungsi untuk mengekstrak fitur penting dari teks, sedangkan *BiLSTM* dipakai untuk memahami makna kalimat dengan menganalisis teks dua arah. Beberapa penelitian membuktikan bahwa model *hybrid* ini dapat menyampaikan hasil yang lebih tepat dibandingkan metode lainnya, seperti *CNN, LSTM, BiLSTM, GRU, Random Forest*, dan *SVM* dalam analisis sentimen berbasis X [10].

Penelitian ini memanfaatkan *BiLSTM* sebagai komponen utama pada model *Hybrid CNN-BiLSTM* dalam metode analisis sentimen atas *tweet* terkait perlindungan data pribadi di Indonesia. Kemampuan *BiLSTM* dalam memahami konteks dari dua arah diharapkan dapat meningkatkan akurasi model dalam mengidentifikasi sentimen yang terkandung dalam teks.

2.2.7 Evaluasi Model

Dalam penelitian analisis sentimen, model dinilai untuk mengetahui seberapa baik dan akurat model tersebut dalam mengelompokkan teks. Metode ini bermaksud untuk mengukur seberapa baik algoritma bisa mengidentifikasi sentimen positif, negatif, dan netral. Pada penelitian yang memakai pendekatan *machine learning* maupun *deep learning*, evaluasi model juga berperan sebagai dasar dalam membandingkan hasil beberapa metode yang dipakai [13].

Dalam penelitian ini, kinerja model *Hybrid CNN-LSTM* dan *Hybrid CNN-BiLSTM* diukur memakai beberapa metrik, yaitu *accuracy, precision, recall*, dan *F1-score.* Akurasi dipakai untuk menilai level kinerja klasifikasi berjalan secara keseluruhan. *Precision* menggambarkan seberapa akurat prediksi model dalam mengenali kelas tertentu, sedangkan *recall* membuktikan tingkat kinerja model dapat mengidentifikasi semua data yang sebenarnya tergolong ke dalam kelas tersebut. *F1-score* ialah metrik yang menggabungkan *precision* dan *recall* untuk membagikan penilaian yang lebih setara mengenai performa model [14].

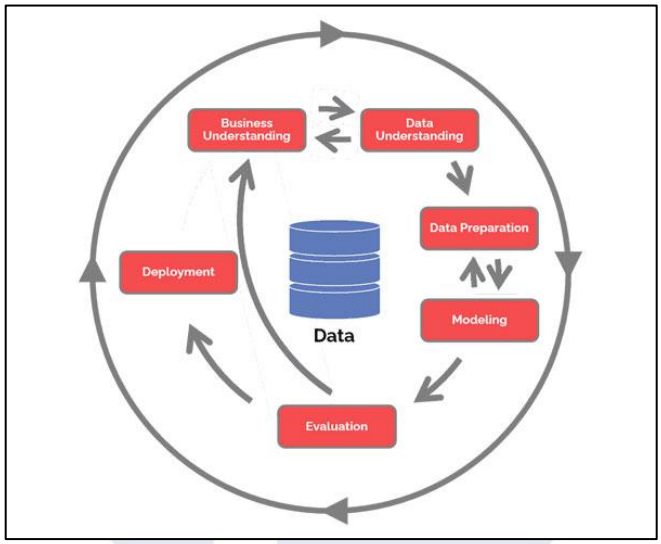
Selain memakai metrik evaluasi seperti *accuracy, precision, recall*, dan *F1-score*, penelitian ini juga memakai *confusion matrix* untuk mendapatkan analisis yang lebih rinci tentang hasil klasifikasi. *Confusion matrix* memudahkan dalam mengetahui berapa ragam prakiraan yang tepat dan tidak tepat untuk setiap kelas sentimen. Dengan demikian, nilai *TP, TN, FP*, dan *FN* dapat dianalisis untuk membagikan pemahaman yang lebih menyeluruh tentang performa model [10].

Pemakaian evaluasi model sangat penting dalam penelitian ini karena hasil pengukuran tersebut menjadi dasar dalam menentukan model terbaik antara *Hybrid CNN-LSTM* dan *Hybrid CNN-BiLSTM*. Model dengan nilai *accuracy, precision, recall,* dan *F1-score* yang lebih tinggi akan dianggap lebih optimal dalam melakukan analisis sentimen atas isu perlindungan data pribadi di Indonesia berdasarkan data X.

2.3 Framework dan Algoritma yang dipakai

2.3.1 CRISP-DM

CRISP-DM ***(****Cross-Industry Standard Process for Data Mining****)*** merupakan metode yang membagikan alur kerja sistematis untuk proyek *data mining*, sehingga metodenya lebih terstruktur, konsisten, dan bisa diterapkan di berbagai bidang industri [19].



Gambar 2. 2 Diagram *Flow CRISP-DM* [20]

Gambar 2.2 membuktikan alur metode *CRISP-DM* yang dipakai dalam penelitian ini. Metode tersebut terdiri dari 6 (enam) tahapan utama yang berperan penting dalam mendukung metode analisis data secara sistematis, mulai dari tahap interpretasi masalah sampai penyampaian hasil penelitian. Adapun uraian dari setiap tahapan dijelaskan sebagai berikut:

1. **Business Understanding**

   Metode ini bertujuan untuk mengetahui permasalahan penelitian, menentukan tujuan yang akan dicapai, serta menetapkan batasan yang dipakai selama penelitian berlangsung. Fokus penelitian diarahkan pada analisis sentimen masyarakat atas isu keamanan data pribadi di Indonesia berdasarkan data X. Selain itu, ditetapkan pula tujuan penelitian berupa perbandingan performa model *Hybrid*

CNN-LSTM dan *Hybrid CNN-BiLSTM* dalam mengelompokkan sentimen positif, negatif, dan netral.

2. **Data Understanding**

   Tahap ini bertujuan untuk memahami karakteristik data yang dipakai dalam penelitian. Data yang dipakai diperoleh melalui metode *crawling* pada *platform* X memakai kata kunci yang berkaitan dengan perlindungan data pribadi. Data dikumpulkan pada periode Januari 2024 hingga Juni 2026, berikutnya dijalankan metode seleksi, pembersihan, dan *preprocessing* sebelum dipakai pada tahap pemodelan. Dataset tersebut berisi kumpulan *tweet* yang berkaitan dengan isu keamanan dan perlindungan data pribadi di Indonesia. Berikutnya dijalankan eksplorasi awal atas data untuk mengetahui jumlah data, distribusi sentimen, serta karakteristik teks seperti pemakaian bahasa informal, singkatan, dan simbol. Tahap ini penting untuk memastikan bahwa data yang dipakai sesuai dengan tujuan penelitian.

3. **Data Preparation**

   Pada tahap ini dijalankan metode persiapan data sebelum dipakai dalam pemodelan. Data terlebih dahulu difilter berdasarkan bahasa Indonesia dan rentang waktu penelitian, berikutnya dipilih atribut yang dipakai dalam metode analisis. Berikutnya dijalankan *preprocessing* yang mencakup perubahan HTML entity, penghapusan *mention, URL*, emoji atau karakter non-ASCII, penghapusan spasi berlebih, serta *case folding*. Setelah itu, dijalankan tokenisasi dan pembentukan kembali teks hasil *preprocessing*. Data berikutnya diberi label memakai pendekatan *lexicon-based labeling*, dikonversi ke bentuk numerik memakai *Label Encoding*, dibagi menjadi data latih dan data uji dengan rasio 80:20, berikutnya diubah menjadi *sequence* memakai *Tokenizer* dan diseragamkan panjangnya memakai padding sebelum dipakai pada tahap pemodelan.

4. **Modeling**

   Tahap modeling dijalankan dengan membangun empat model, yaitu *CNN, LSTM, Hybrid CNN-LSTM,* dan *Hybrid CNN-BiLSTM. CNN* dan *LSTM* dipakai sebagai model pembanding (*baseline*), sedangkan *Hybrid CNN-LSTM* dan *Hybrid CNN-BiLSTM* dipakai sebagai model utama dalam penelitian. Data yang telah dipersiapkan pada tahap sebelumnya dipakai sebagai data masukan untuk kedua model. *CNN* berperan sebagai *feature extractor,* sedangkan *LSTM* dan *BiLSTM* dipakai untuk memahami konteks serta keterkaitan antar kata dalam teks. Melalui kombinasi tersebut, model diharapkan mampu memberikan performa klasifikasi sentimen yang lebih baik dibandingkan dengan pemakaian metode tunggal.

5. **Evaluation**

   Evaluasi model dijalankan setelah metode pelatihan selesai untuk mengetahui tingkat keberhasilan model dalam mengelompokkan sentimen. Pengukuran performa memakai metrik *accuracy, precision, recall,* dan *F1-score,* sedangkan *confusion matrix* dipakai untuk menganalisis hasil klasifikasi secara lebih detail pada setiap kelas sentimen. Hasil yang diperoleh dari model *Hybrid CNN-LSTM* dan *Hybrid CNN-BiLSTM* berikutnya dibandingkan sebagai dasar dalam menentukan model yang paling optimal untuk analisis sentimen data X. *CNN* dan *LSTM* juga dievaluasi sebagai model *baseline* sehingga performanya dapat dibandingkan dengan kedua model *hybrid*.

6. **Deployment**

   Tahap *deployment* dalam metode *CRISP-DM* bertujuan untuk menerapkan model yang sudah dibuat agar bisa dipakai dalam sistem atau aplikasi. Pada tahap *deployment,* model yang memiliki hasil terbaik diterapkan ke dalam aplikasi berbasis *Streamlit,* sehingga pemakai bisa melakukan prediksi sentimen atas teks yang membahas perlindungan data pribadi secara interaktif. Implementasi

ini bertujuan untuk membuktikan penerapan model hasil penelitian pada suatu aplikasi sederhana berbasis web.

2.3.2 Hybrid CNN-LSTM

Hybrid CNN-LSTM adalah model *deep learning* yang dibangun melalui kombinasi *CNN* dan *LSTM* dalam satu *framework*, sehingga mampu memanfaatkan keunggulan kedua metode dalam pengolahan teks. Pada model ini, *CNN* berfungsi sebagai *feature extractor* yang dapat mengenali pola lokal, seperti kata atau frasa penting dalam teks, sedangkan *LSTM* dipakai untuk memahami hubungan antar kata berdasarkan urutan dalam kalimat. Dalam metodenya, *CNN* terlebih dahulu melakukan konvolusi untuk mengambil fitur penting dari data teks. Setelah itu, hasil ekstraksi fitur diteruskan ke *LSTM* untuk mempelajari hubungan konteks antar kata. Melalui kombinasi tersebut, model mampu memahami pola kata sekaligus konteks urutan kata dalam suatu kalimat [7].

Penggabungan *CNN* dan *LSTM* memberikan keunggulan dibandingkan dengan model tunggal karena mampu mengombinasikan metode ekstraksi fitur dan pemodelan sekuensial secara bersamaan [21]. Pada data teks tidak teratur, seperti yang berasal dari media sosial, *Hybrid CNN-LSTM* mampu memberikan performa yang baik karena dapat mempelajari pola lokal serta informasi kontekstual secara bersamaan dalam metode klasifikasi.

Pada penelitian ini, model *Hybrid CNN-LSTM* dipakai sebagai salah satu metode untuk melakukan klasifikasi sentimen pada data X terkait isu perlindungan data pribadi di Indonesia. Model tersebut berikutnya dibandingkan dengan *Hybrid CNN-BiLSTM* untuk memahami model yang memiliki kinerja lebih baik dalam metode analisis sentimen.

2.3.3 Hybrid CNN-BiLSTM

Dalam arsitektur *Hybrid CNN-BiLSTM*, *CNN* dipakai untuk mengekstrak fitur yang bisa mengenali pola penting dalam teks, seperti kata dan frasa tertentu. Berikutnya, *BiLSTM* dipakai untuk mengkaji hubungan antar kata

dengan meninjau konteks dari dua arah sekaligus. Menggabungkan kedua metode itu bertujuan agar model lebih mampu memahami dan mengelompokkan data teks [5].

Pada model ini, *CNN* terlebih dahulu melakukan metode konvolusi untuk mengekstraksi fitur dari data teks. Hasil ekstraksi tersebut berikutnya dilanjutkan ke *BiLSTM* untuk mempelajari hubungan antar kata tidak hanya berdasarkan urutan sebelumnya, tetapi juga mempertimbangkan kata setelahnya dalam kalimat. Dengan cara ini, model dapat memahami konteks kalimat secara lebih lengkap.

Pemakaian *BiLSTM* memungkinkan model untuk menangkap informasi yang lebih luas dibandingkan *LSTM* biasa karena mempertimbangkan dua arah dalam metode pemrosesan data teks [6]. Kombinasi *CNN* dan *BiLSTM* memungkinkan model untuk menggabungkan metode ekstraksi fitur dan pemahaman konteks dua arah secara bersamaan sehingga menciptakan representasi teks yang lebih baik dalam klasifikasi sentimen [10].

Penelitian ini menerapkan *Hybrid CNN-BiLSTM* sebagai model pembanding bagi *Hybrid CNN-LSTM* guna mengidentifikasi model yang memiliki kinerja lebih optimal dalam mengelompokkan sentimen pada data X terkait perlindungan data pribadi di Indonesia.

2.4 Tools dan software yang dipakai

2.4.1 Python

Python ialah salah satu bahasa pemrograman yang sering dipakai dalam analisis data. *Python* banyak dipakai untuk membantu metode *data mining*, *machine learning*, serta *deep learning*. Kemudahan sintaks serta ketersediaan berbagai *library* menjadikan *Python* efektif dipakai dalam pengolahan data dan pengembangan model. Pada penelitian ini, *Python* dimanfaatkan untuk menjalankan metode *preprocessing*, pelatihan model, dan evaluasi hasil dengan bantuan *library* seperti *NumPy*, pandas, *scikit-learn*, dan *TensorFlow* [22]. Selain itu, *Python* juga mampu mendukung integrasi berbagai tahapan analisis

secara efisien sehingga dipakai sebagai *platform* utama dalam pembangunan model *CNN, LSTM, Hybrid CNN-LSTM* dan *Hybrid CNN-BiLSTM* [23]. Oleh sebab itu, penelitian ini memakai *Python* sebagai bahasa pemrograman utama untuk melakukan *preprocessing* data, membangun model *CNN*, *LSTM, Hybrid CNN-LSTM*, dan *Hybrid CNN-BiLSTM*, serta melakukan evaluasi hasil analisis sentimen.

2.4.2 Jupyter Notebook

Dalam bidang analisis data dan *machine learning*, *Jupyter Notebook* banyak dipakai sebagai lingkungan kerja berbasis web yang mendukung eksekusi kode secara interaktif [24]. Selain menjalankan kode program, *platform* ini juga memungkinkan penyajian teks dan visualisasi dalam satu dokumen yang terintegrasi, sehingga metode eksplorasi data dan pembangunan model dapat dijalankan dengan lebih terstruktur dan efisien.

Pemakaian *Jupyter Notebook* dalam analisis data dan *machine learning* juga memungkinkan integrasi antara kode, visualisasi, dan dokumentasi dalam satu platform, sehingga dapat mendukung metode eksplorasi data serta *reproducibility* dalam penelitian [24]. Selain itu, pemakaian *Jupyter Notebook* dapat membantu meningkatkan pemahaman pemakai atas alur analisis karena hasil dan metode dapat ditampilkan secara langsung dalam satu tampilan [25].

Sebagai lingkungan pengembangan utama, *Jupyter Notebook* dipakai dalam penelitian ini untuk menjalankan *preprocessing data*, pembangunan model *CNN, LSTM, Hybrid CNN-LSTM*, dan *Hybrid CNN-BiLSTM*, serta evaluasi hasil klasifikasi sentimen.

2.4.3 Streamlit

Streamlit merupakan *framework open-source* yang memakai bahasa *Python dan* dipakai untuk membentujk aplikasi web interaktif dengan cara yang cepat dan mudah, terutama untuk aplikasi yang berhubungan dengan *data science, machine learning,* dan *artificial intelligence*. *Streamlit* memudahkan pengembang untuk mengubah kode *Python* menjadi aplikasi web tanpa

memerlukan pengetahuan mendalam mengenai HTML, CSS, maupun JavaScript. Sehingga, metode membuat dan menerapkan model *machine learning* menjadi lebih cepat dan lebih sederhana. Selain mudah diimplementasikan, *Streamlit* juga memiliki berbagai komponen interaktif seperti tombol (*button*), kotak teks (*text input*), *slider*, *checkbox*, serta kemampuan untuk menampilkan visualisasi data yang bisa langsung dipakai dengan *library Python*, seperti *Pandas, Matplotlib,* dan *Plotly*. *Framework* ini juga memiliki fitur *live reload*, sehingga setiap kali ada perubahan pada kode, tampilan aplikasi akan langsung diperbarui tanpa perlu melakukan konfigurasi ulang. Ciri khas tersebut membuat *Streamlit* sebagai salah satu *framework* yang banyak dipakai untuk membangun aplikasi berbasis *machine learning* secara cepat dan interaktif [26].

Pada penelitian ini, Streamlit dipakai sebagai alat *deployment* model *CNN, LSTM, Hybrid CNN-LSTM,* dan *Hybrid CNN-BiLSTM*. Setelah metode pelatihan dan evaluasi selesai, keempat model digabungkan ke dalam aplikasi berbasis *Streamlit*, sehingga pemakai bisa memilih model yang akan dipakai untuk memprediksi sentimen teks terkait perlindungan data pribadi melalui antarmuka web yang mudah dipakai dan sederhana.

2.4.4 Visual Studio Code

Visual Studio Code ialah editor kode sumber (*source-code editor*) yang dibangun oleh *Microsoft* dan sering dipakai dalam membuat perangkat lunak, termasuk aplikasi berbasis *Python*. *Visual Studio Code* bisa dipakai di berbagai jenis sistem operasi dan menawarkan lingkungan kerja pengembangan yang ringan, mudah dipakai, serta bisa diperluas dengan ekstensi-ekstensi yang sesuai dengan kebutuhan pemakai. Dengan memakai berbagai ekstensi, *Visual Studio Code* bisa berperan sebagai *Integrated Development Environment (IDE)* yang membantu metode membuat aplikasi dengan lebih efisien. Selain tampilan yang mudah dipakai, *Visual Studio Code* memiliki fitur seperti *IntelliSense, syntax highlighting, debugging*, terminal terintegrasi, serta dukungan berbagai ekstensi yang membantu metode penulisan, pengujian, dan pengelolaan kode

program. Pemakaian *Visual Studio Code* juga dinilai mampu meningkatkan pengalaman pemrograman karena mudah dipakai, memiliki fungsionalitas yang lengkap, serta banyak dipakai oleh pengembang perangkat lunak [27].

Pada penelitian ini, *Visual Studio Code* dipakai sebagai alat elaborasi untuk membuat aplikasi yang memakai *Streamlit*. Perangkat lunak ini dipakai untuk membuat, menggabungkan, serta menguji aplikasi yang memakai model *CNN, LSTM, Hybrid CNN-LSTM,* dan *Hybrid CNN-BiLSTM* sehingga metode *deployment* model analisis sentimen bisa dijalankan secara lebih terstruktur dan efisien.