

BAB I

PENDAHULUAN

1.1 Latar Belakang

Fenomena keamanan data pribadi di Indonesia semakin banyak mendapat perhatian karena meningkatnya kasus kebocoran data pada era digital. Selama beberapa tahun terakhir, berbagai peristiwa kebocoran data yang melibatkan masyarakat Indonesia terjadi pada beberapa *platform* digital, seperti layanan *e-commerce* dan instansi pemerintah. Perkembangan teknologi yang semakin pesat menyebabkan metode digitalisasi data terus berkembang sehingga diperlukan upaya perlindungan data yang lebih baik. Menurut Undang-Undang Nomor 23 Tahun 2006 tentang Administrasi Kependudukan Pasal 1 ayat 22, data pribadi didefinisikan sebagai data individu yang disimpan, dipelihara, serta dijaga keakuratannya, dan keberadaannya dilindungi oleh negara [1]. Selain meningkatnya kasus kebocoran data, pemerintah Indonesia juga mengesahkan Undang-Undang Nomor 27 Tahun 2022 tentang Perlindungan Data Pribadi (UU PDP) sebagai dasar hukum dalam menjaga keamanan data masyarakat. Regulasi itu menetapkan hak bagi pemilik data, kewajiban untuk pengelola data pribadi, dan penerapan sanksi atas pihak yang tidak mematuhi peraturan perlindungan data pribadi [2]. Selama beberapa tahun terakhir, industri teknologi informasi dan komunikasi di Indonesia mengalami pertumbuhan yang terus berlanjut dan pemakaiannya semakin luas. Pemakaian internet di Indonesia semakin maju dan sudah dimanfaatkan oleh penduduk di berbagai daerah, baik yang ada di kota maupun yang di desa. Berdasarkan laporan Statistik Telekomunikasi Indonesia 2024 dari Badan Pusat Statistik (BPS), jumlah orang yang mengakses internet di Indonesia terus bertambah sejak tahun 2021 sampai tahun 2024 [3].



Gambar 1. 1 Persentase Penduduk yang Mengakses Internet Menurut Klasifikasi Daerah, 2021-2024 [3]

Berdasarkan Gambar 1.1, pemakaian internet di Indonesia membuktikan pola pertumbuhan yang stabil setiap tahunnya. Persentase pengguna internet tercatat sebanyak 62,10% pada tahun 2021, berikutnya meningkat menjadi 66,48% pada tahun 2022, 69,21% pada tahun 2023, dan mencapai 72,78% pada tahun 2024. Tren tersebut mengindikasikan bahwa pemanfaatan internet oleh masyarakat Indonesia semakin luas dalam berbagai aktivitas sehari-hari, sehingga tingkat keterlibatan dan konektivitas masyarakat atas ekosistem digital juga terus meningkat [3].

Seiring dengan meningkatnya akses internet, pemakaian media sosial di Indonesia juga terus mengalami perkembangan. Media sosial saat ini banyak dimanfaatkan masyarakat sebagai sarana komunikasi, berbagi informasi, dan menyampaikan pendapat terkait berbagai isu yang sedang berkembang. Sebagai salah satu jaringan media sosial yang terkenal, X (sebelumnya dikenal dengan nama *Twitter*) banyak dimanfaatkan oleh orang-orang untuk mengungkapkan pandangan, memberikan kritik, serta menanggapi berbagai masalah secara transparan. Platform tersebut memungkinkan pengguna untuk membagikan pendapat secara singkat dan cepat melalui fitur *tweet* sehingga banyak dipakai dalam pembahasan berbagai topik, termasuk isu keamanan data pribadi. Adanya

berbagai opini yang diungkapkan pemakai, baik yang bernilai positif, negatif, maupun netral, menjadikan X sering dipakai sebagai referensi data dalam studi untuk menggambarkan pandangan masyarakat atas suatu isu [4], [5], [6].

Dalam studi ini, data didapat melalui metode *crawling* di *platform* X dengan istilah yang berkaitan dengan keamanan data pribadi. Data yang dipakai mencakup rentang waktu dari Januari 2024 hingga Juni 2026 sehingga mampu merepresentasikan opini masyarakat atas isu perlindungan data pribadi berdasarkan data yang lebih mutakhir dibandingkan penelitian-penelitian sebelumnya. Dalam pengolahan data opini yang jumlahnya terus meningkat, analisis sentimen sering dipakai untuk mendapatkan informasi mengenai kecenderungan pandangan masyarakat. Analisis sentimen adalah metode dalam bidang *Natural Language Processing (NLP)* yang memiliki tujuan guna mengenali dan mengkategorikan isi teks sesuai sentimen yang terdapat di dalamnya, seperti sentimen positif, negatif, maupun netral. Metode yang disebutkan sebelumnya telah dipakai secara luas dalam berbagai studi untuk memahami sudut pandang banyak orang mengenai suatu permasalahan di berbagai sektor, termasuk sosial, ekonomi, kesehatan, dan teknologi [4], [7]. Analisis sentimen berkontribusi dalam mengonversi data yang awalnya tidak teratur dalam format tertulis menjadi informasi yang lebih tersusun, sehingga mempermudah para peneliti dalam melaksanakan analisis dan pengambilan kesimpulan.

Kemajuan teknologi pengolahan bahasa alami telah mendorong pemakaian *deep learning* dan *hybrid deep learning* dalam analisis sentimen. Dibandingkan dengan metode machine learning tradisional, dalam pendekatan ini mempunyai kualifikasi yang lebih baik dalam memahami pola pada data teks berukuran besar. Salah satu pendekatan *deep learning* yang banyak dipakai untuk tugas tersebut adalah *Long Short-Term Memory (LSTM)*. Metode ini mampu mengolah teks secara berurutan sehingga dapat menafsirkan ikatan antar kata dan konteks dalam kalimat dengan lebih efektif. *Long Short-Term Memory (LSTM)* merupakan evolusi dari *RNN* yang dibuat agar mampu menyimpan

serta memahami data dalam periode waktu yang lebih lama. Kemampuan ini membuat *LSTM* sering dipakai dalam studi analisis sentimen, terutama bagi data dari media sosial yang mempunyai ciri-ciri teks yang kompleks dan berurutan [8].

Salah satu evolusi dari *LSTM* adalah *Bidirectional Long Short-Term Memory (BiLSTM)*. Berbeda dengan *LSTM* biasa, *BiLSTM* dapat mengetahui konteks kalimat dengan menganalisis urutan kata dari dua arah, yakni dari awal ke akhir (*forward*) dan dari akhir ke awal (*backward*). Melalui mekanisme pemrosesan dua arah, *BiLSTM* mampu memahami arti dan konteks kalimat dengan cara yang lebih menyeluruh, sehingga kinerjanya dalam tugas klasifikasi teks umumnya lebih unggul dibandingkan dengan *LSTM* standar. Berdasarkan penelitian sebelumnya, metode *BiLSTM* mendapatkan tingkat akurasi sebanyak 91,4% dan dinilai lebih efektif dibandingkan dengan *Recurrent Neural Network (RNN)* dalam metode klasifikasi sentimen. Hal tersebut membuktikan bahwa *BiLSTM* lebih mampu memahami struktur bahasa alami yang kompleks [9].

Perkembangan analisis sentimen sekarang ini tidak cuma terpusat pada penerapan satu tipe model *deep learning*, melainkan juga pada model *hybrid* yang menggabungkan *Convolutional Neural Network (CNN)* dengan *LSTM* atau *BiLSTM* guna mendapatkan hasil klasifikasi yang lebih optimal. Pada pendekatan ini, *CNN* dimanfaatkan untuk mendapatkan fitur-fitur penting dari teks, seperti pola kata dan frasa yang sering muncul. Sedangkan *LSTM* dan *BiLSTM* dipakai untuk menganalisis konteks serta keterkaitan kata dalam kalimat sehingga interpretasi atas konteks teks menjadi lebih optimal. Penggabungan kedua metode tersebut dapat meningkatkan hasil klasifikasi sentimen karena *CNN* berfungsi dalam metode ekstraksi fitur, sedangkan *LSTM* dan *BiLSTM* membantu memahami konteks kalimat dengan lebih baik [7], [10].

Penelitian terdahulu membuktikan model *Hybrid CNN-BiLSTM* mendapatkan tingkat akurasi tertinggi sebanyak 96,59% jika dibandingkan dengan beberapa metode lain, seperti *CNN*, *LSTM*, *BiLSTM*, *GRU*, *SVM*, dan

Random Forest [10]. Berdasarkan hasil penelitian sebelumnya, kombinasi *CNN* dan *BiLSTM* berpotensi menghasilkan performa yang sangat baik pada analisis sentimen X. Sedangkan, model *Hybrid CNN-LSTM* juga membuktikan tingkat akurasi yang tinggi, yaitu sebanyak 94% [7]. Hasil tersebut mengindikasikan bahwa metode hybrid kemampuan untuk menghasilkan kinerja yang lebih baik dibandingkan dengan model tunggal karena dapat menggabungkan keunggulan dari masing-masing teknik yang diterapkan.

Merujuk pada dampak penelitian yang telah dijalankan sebelumnya, diperlukan analisis perbandingan antara model *Hybrid CNN-LSTM* dan *Hybrid CNN-BiLSTM* untuk mengidentifikasi model yang berpotensi menghasilkan performa klasifikasi sentimen yang lebih baik. Perbandingan kedua model tersebut penting dijalankan karena keduanya memakai *CNN* sebagai metode ekstraksi fitur, tetapi memiliki perbedaan pada cara memahami konteks kalimat, yaitu melalui *LSTM* dan *BiLSTM*. Dengan demikian, penelitian ini memanfaatkan kedua model tersebut untuk mengetahui model *hybrid* yang memberikan performa terbaik dalam mengelompokkan sentimen positif, negatif, dan netral pada data media sosial X terkait perlindungan data pribadi.

1.2 Rumusan Masalah

Dari latar belakang yang telah dipaparkan, oleh karena itu rumusan masalah yang terdapat dalam penelitian ini adalah sebagai berikut:

1. Bagaimana hasil kinerja model *Hybrid CNN-LSTM* dalam mengelompokkan sentimen masyarakat atas isu keamanan data pribadi di Indonesia?
2. Bagaimana hasil kinerja model *Hybrid CNN-BiLSTM* dalam mengelompokkan sentimen masyarakat atas isu keamanan data pribadi di Indonesia?
3. Model manakah yang memiliki performa terbaik antara *Hybrid CNN-LSTM* dan *Hybrid CNN-BiLSTM* dalam analisis sentimen data X?

1.3 Batasan Masalah

Agar penelitian ini mencapai hasil yang lebih fokus, batasan masalah yang dijadikan acuan dalam penelitian ini adalah sebagai berikut:

1. Data yang dipakai berupa *tweet* pada *platform* X yang membahas isu perlindungan data pribadi di Indonesia dan dikumpulkan pada periode Januari 2024 hingga Juni 2026.
2. Analisis sentimen dibatasi pada tiga kategori, yaitu positif, negatif, dan netral.
3. Metode yang dipakai dalam penelitian ini mencakup *CNN*, *LSTM*, *Hybrid CNN-LSTM* dan *Hybrid CNN-BiLSTM* dengan *CNN* dan *LSTM* dipakai sebagai model pembandingan (*baseline*).
4. Evaluasi kinerja model dijalankan memakai metrik *accuracy*, *precision*, *recall*, *F1-score*, dan *confusion matrix*.
5. Penelitian ini mencakup tahap *deployment* dalam bentuk aplikasi berbasis *Streamlit* untuk mengimplementasikan model *CNN*, *LSTM*, *Hybrid CNN-LSTM*, dan *Hybrid CNN-BiLSTM* sehingga memungkinkan pemakai untuk melakukan prediksi sentimen secara langsung.

1.4 Tujuan dan Manfaat Penelitian

1.4.1 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Memahami performa model *Hybrid CNN-LSTM* dalam metode klasifikasi sentimen masyarakat terkait isu keamanan data pribadi di Indonesia.
2. Mengetahui kinerja model *Hybrid CNN-BiLSTM* dalam metode klasifikasi sentimen masyarakat terkait isu keamanan data pribadi di Indonesia.
3. Membandingkan performa model *Hybrid CNN-LSTM* dan *Hybrid CNN-BiLSTM* untuk menentukan model yang paling optimal dalam analisis sentimen data X.

4. Mengaplikasikan hasil penelitian ke dalam aplikasi berbasis *Streamlit* sebagai media *deployment* untuk melakukan prediksi sentimen atas teks mengenai perlindungan data pribadi.

1.4.2 Manfaat Penelitian

Melalui Penelitian ini ditujukan guna dapat memberi manfaat sebagai berikut:

1. Penelitian ini diharapkan memberikan manfaat akademis berupa kontribusi atas pengembangan ilmu Sistem Informasi, terutama dalam bidang *NLP*, *Text Mining*, dan Analisis Sentimen. Terlebih lagi, penelitian ini mampu menjadi acuan tambahan terkait perbandingan model *Hybrid CNN-LSTM* dan *Hybrid CNN-BiLSTM* dengan *CNN* dan *LSTM* sebagai *baseline* untuk menganalisis sentimen masyarakat pada platform X.
2. Manfaat Praktis memberikan gambaran kepada pemerintah, lembaga keamanan siber, dan penyelenggara sistem elektronik mengenai persepsi serta sentimen masyarakat atas penerapan Undang-Undang Perlindungan Data Pribadi (UU PDP). Hasil penelitian bisa dipakai selaku bahan penilaian dalam cara evaluasi prosedur dan pengembangan strategi perlindungan data pribadi yang lebih efektif. Berikutnya, hasil penelitian direalisasikan dalam bentuk aplikasi berbasis *Streamlit* sehingga dapat dipakai sebagai media demonstrasi dan prediksi sentimen atas teks mengenai perlindungan data pribadi.
3. Manfaat bagi pemakai membantu masyarakat dalam memahami kecenderungan opini publik terkait isu keamanan data pribadi, sehingga dapat mengembangkan pemahaman akan pentingnya perlindungan data serta mendorong sikap yang lebih kritis atas kebijakan dan insiden kebocoran data.
4. Manfaat analitis dari penelitian ini adalah memberikan gambaran mengenai perbandingan performa model *Hybrid*

CNN-LSTM dan *Hybrid CNN-BiLSTM* dengan dukungan model *CNN* dan *LSTM* sebagai *baseline*, sehingga dapat diidentifikasi model yang paling efektif dalam mengelompokkan data teks media sosial pada kategori positif, negatif, dan netral.

1.5 Sistematika Penulisan

Adapun sistematika penulisan yang dipakai dalam penelitian ini adalah sebagai berikut:

BAB I PENDAHULUAN

Bab ini menguraikan dasar-dasar dalam melakukan penelitian, mulai dari latar belakang penelitian, perumusan masalah, batasan-batasan penelitian, tujuan penelitian dan manfaat yang diperoleh, serta struktur penulisan yang akan diikuti. Selain itu, dijelaskan juga gambaran umum mengenai analisis sentimen atas isu perlindungan data pribadi yang menjadi fokus penelitian.

BAB II LANDASAN TEORI

Bab ini menguraikan berbagai gagasan yang dipakai sebagai asas dalam penelitian, yaitu konsep tentang keamanan dan perlindungan data pribadi, analisis sentimen, *Natural Language Processing (NLP)*, *Convolutional Neural Network (CNN)*, *Long Short-Term Memory (LSTM)*, *Bidirectional Long Short-Term Memory (BiLSTM)*, serta model *Hybrid CNN-LSTM* dan *Hybrid CNN-BiLSTM*. Di samping itu, bab ini juga membahas penelitian sebelumnya yang penting sebagai acuan dan bahan untuk membandingkan dalam penelitian.

BAB III METODOLOGI PENELITIAN

Bab ini menguraikan cara penelitian dilaksanakan, yaitu mulai dari mengumpulkan data dengan melakukan *crawling* di *platform X*, *preprocessing data*, pelabelan sentimen, pembangunan model *CNN*, *LSTM*, *Hybrid CNN-LSTM*, dan *Hybrid CNN-BiLSTM*, metode pelatihan model, evaluasi memakai metrik *accuracy*, *precision*, *recall*, *F1-score*, dan *confusion matrix*, serta tahap *deployment* model memakai aplikasi berbasis *Streamlit*.

BAB IV ANALISIS DAN HASIL PEMBAHASAN

Bab ini menguraikan hasil dari penerapan penelitian, mencakup hasil *preprocessing*, pelabelan sentimen, evaluasi performa model *CNN*, *LSTM*, *Hybrid CNN-LSTM*, dan *Hybrid CNN-BiLSTM*, serta perbandingan hasil kinerja masing-masing model. Selain itu, juga dijelaskan bagaimana model tersebut diimplementasikan dalam aplikasi berbasis *Streamlit* sebagai media *deployment* untuk prediksi sentimen.

BAB V SIMPULAN DAN SARAN

Bab ini berisi kesimpulan dari hasil penelitian yang sudah dijalankan dengan memakai berbagai metode, serta beberapa saran yang bisa dimanfaatkan dalam pengembangan penelitian berikutnya, baik dalam hal metode, algoritma, maupun pengembangan sistem analisis sentimen.

