Lentera: Multidisciplinary Studies
Volume 3 Number 2, March, 2025
p- ISSN: 2987-2472 | e-ISSN: 2897-7031

ANALYSIS OF *DROPOUT* AND LEARNING RATE ON BILSTM-ANN PERFORMANCE IN HATE SPEECH DETECTION

Ranny Erzitha, Syafrijon, Dony Novaliendry, Khairi Budayawan
Universitas Negeri Padang, Padang, Indonesia
rannyerzithal@student.unp.ac.id

ABSTRAK

Penyebaran ujaran kebencian di media sosial semakin meningkat, sehingga menuntut pengembangan model deteksi yang akurat dan efisien. Penelitian ini menganalisis dampak dari parameter dropout dan learning rate terhadap kinerja model BiLSTM-ANN dalam mendeteksi ujaran kebencian di media sosial Indonesia. Tujuan utama dari penelitian ini adalah untuk mengevaluasi akurasi, F1-score, precision, recall, dan AUC yang dihasilkan oleh model, serta untuk menentukan nilai optimal untuk learning rate dan dropout. Model diuji menggunakan data komentar dari platform media sosial Indonesia. Metode penelitian menggunakan pendekatan CRISP-DM, dengan pengumpulan 2064 komentar dari TikTok dan Twitter (X), yang kemudian diproses melalui berbagai tahap, termasuk prapemrosesan teks, pembelajaran mendalam dengan BiLSTM-ANN, serta evaluasi performa model. Model diuji dengan berbagai kombinasi dropout (0,02–0,7) dan learning rate (0,00001–0,01) untuk menemukan konfigurasi optimal. **Hasil uji** menunjukkan bahwa model BiLSTM-ANN mencapai akurasi sebesar 68,12%, dengan precision sebesar 53,71% dan recall sebesar 82,55%. Meskipun recall model cukup tinggi, precision yang relatif rendah menunjukkan kesulitan dalam mendeteksi kelas minoritas. Ketidakseimbangan data dan variasi bahasa di media sosial menjadi tantangan utama bagi model ini. Penelitian ini menyimpulkan bahwa optimasi parameter, peningkatan data pelatihan, dan penerapan teknik yang lebih canggih diperlukan untuk meningkatkan kinerja deteksi ujaran kebencian, dengan implikasi untuk sistem moderasi konten di platform media sosial. Implikasi penelitian ini mencakup perbaikan dalam sistem moderasi konten berbasis kecerdasan buatan, dengan potensi implementasi model BiLSTM-ANN pada platform media sosial untuk meningkatkan efektivitas dalam mengidentifikasi ujaran kebencian secara real-time. Diperlukan penelitian lanjutan dengan dataset yang lebih besar serta penerapan teknik balancing data untuk meningkatkan akurasi dan generalisasi model.

Kata Kunci: *BiLSTM-AN, droupout,* tingkat pembelajaran, deteksi ujaran kebencian, media sosial

Abstract

The spread of hate speech on social media is increasing, thus demanding the development of accurate and efficient detection models. This study analyzes the impact of dropout parameters and learning rate on the performance of the BiLSTM-ANN model in detecting hate speech on Indonesian social media. The main purpose of this study is to evaluate the accuracy, F1-score, precision, recall, and AUC generated by the model, as well as to determine the optimal values for learning rate and dropout. The model was tested using comment data from Indonesian social media platforms. The research method used the CRISP-DM approach, with the collection of 2064 comments from TikTok and Twitter (X), which were then processed through various stages, including text preprocessing, deep learning with BiLSTM-ANN, and model performance evaluation. The model was tested with various combinations of dropout (0.02–0.7) and learning rate (0.00001–0.01) to find the optimal configuration. The test results show that the BiLSTM-ANN model achieves an accuracy of 68.12%, with a precision of 53.71% and a recall of 82.55%. Although the model recall is quite high, the relatively low precision indicates difficulty in detecting minority classes. Data imbalances and language variations on social media are the main challenges for this model. The study concludes that parameter optimization, improved training data, and the application of more sophisticated techniques are needed to improve hate speech detection performance, with implications for content moderation systems on social media platforms. The implications of this study include improvements in artificial intelligence-based content moderation systems, with the potential implementation of the BiLSTM-ANN model on social media platforms to improve effectiveness in identifying hate speech in real-time. Further research with larger datasets and

the application of data balancing techniques are needed to improve the accuracy and generalization of the model.

***Keywords**: BiLSTM-ANN, dropout, learning rate, hate speech detection, social media*

PENDAHULUAN

Analisis sentimen merupakan cabang dari *Natural Language Processing* (NLP) yang berfokus pada identifikasi dan pengkategorian emosi, opini, dan sikap yang diekspresikan dalam teks. Seiring dengan semakin populernya media sosial, analisis sentimen menjadi sangat penting, terutama dalam deteksi ujaran kebencian. Platform seperti Twitter (X), Facebook, dan TikTok menghasilkan data dalam jumlah besar setiap harinya, yang menciptakan kebutuhan mendesak untuk analisis konten, khususnya dalam mendeteksi ujaran kebencian (Abdrakhmanov et al., 2024).

Penyebaran ujaran kebencian tidak hanya mempengaruhi individu, tetapi juga memicu konflik sosial dan merusak kohesi masyarakat (Toktarova et al., 2023).

Meskipun analisis sentimen mencakup berbagai emosi, deteksi ujaran kebencian adalah subkategori yang lebih spesifik dan menantang. Hal ini disebabkan oleh variasi bahasa yang digunakan dan ketidakseimbangan data, yang memerlukan penggunaan model machine learning yang lebih canggih. Ujaran kebencian dapat mengambil berbagai bentuk, seperti ekspresi kebencian terhadap ras, agama, gender, atau kelompok tertentu. Oleh karena itu, deteksi ujaran kebencian memerlukan model yang mampu memahami nuansa bahasa yang kompleks dan konteks sosial.

Dampak dari ujaran kebencian sangat luas, baik dalam aspek individu maupun sosial. Secara individual, ujaran kebencian dapat menyebabkan gangguan psikologis, seperti kecemasan dan depresi bagi korban. Secara sosial, ujaran kebencian dapat memperparah polarisasi kelompok, memicu kekerasan, dan memperburuk stabilitas sosial serta politik. Oleh karena itu, pengembangan sistem deteksi ujaran kebencian yang lebih akurat dan efektif menjadi kebutuhan mendesak.

Beberapa penelitian di Indonesia telah fokus pada deteksi Ujaran Kebencian dan Bahasa Kasar (HSAL), dengan memanfaatkan model seperti LSTM dan BiLSTM. Penelitian sebelumnya menunjukkan bahwa penggunaan embeddings GloVe dengan model klasifikasi LSTM meningkatkan akurasi dalam deteksi ujaran kebencian, meskipun tantangan seperti overfitting akibat data yang tidak seimbang tetap ada (Hayaty, Laksito, & Adi, 2023). Model deep learning, khususnya BiLSTM-ANN, telah terbukti efektif dalam analisis sentimen (Puttarattanamanee, Boongasame, & Thammarak, 2023), tetapi kinerjanya sangat bergantung pada pemilihan parameter yang tepat, khususnya dropout dan learning rate (Af'idah, Anggraeni, Rizki, Setiawan, & Handayani, 2023), (Setiawan & Lestari, 2021)

Meskipun dropout dan learning rate mempengaruhi kinerja model secara signifikan, temuan penelitian beragam. Beberapa penelitian menunjukkan bahwa tingkat dropout yang lebih tinggi meningkatkan kinerja model, sementara yang lain menunjukkan sebaliknya (Hesaputra, 2023; Setiawan & Lestari, 2021). Selain itu, learning rate yang tidak tepat dapat menghambat konvergensi model (Insani, 2023). Oleh karena itu,

eksperimen dengan berbagai kombinasi parameter ini sangat penting untuk menemukan konfigurasi yang optimal.

Penelitian ini berfokus pada pengembangan model BiLSTM-ANN untuk mendeteksi ujaran kebencian di media sosial Indonesia. Model ini memanfaatkan kombinasi Long Short-Term Memory (LSTM) dengan jaringan saraf tiruan (ANN) untuk meningkatkan akurasi klasifikasi teks. Variabel utama yang dikaji dalam penelitian ini adalah dropout dan learning rate, dua parameter penting yang dapat mempengaruhi kinerja model dalam mendeteksi ujaran kebencian.

Kebaruan dari penelitian ini terletak pada eksplorasi mendalam terhadap pengaruh parameter dropout dan learning rate dalam meningkatkan performa model BiLSTM-ANN. Meskipun berbagai studi telah mengkaji deteksi ujaran kebencian dengan menggunakan model machine learning, penelitian ini memberikan kontribusi baru dengan menyesuaikan konfigurasi parameter untuk mengatasi permasalahan ketidakseimbangan data dan variasi bahasa di media sosial Indonesia.

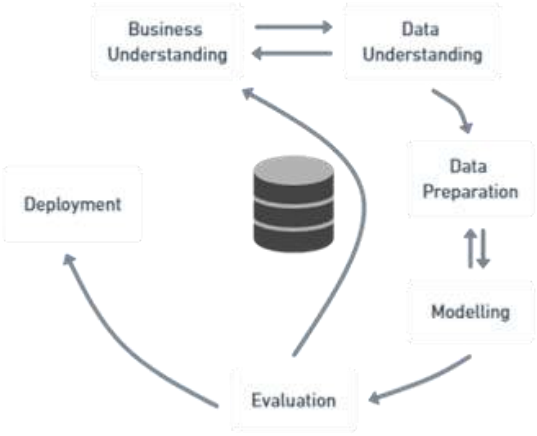
Urgensi penelitian ini didasarkan pada meningkatnya kebutuhan akan sistem moderasi konten otomatis yang lebih andal, mengingat volume besar data yang dihasilkan di media sosial setiap harinya. Sistem deteksi yang lebih akurat dapat membantu platform media sosial dalam mengidentifikasi dan menghapus konten berbahaya sebelum menyebar lebih luas.

Penelitian ini bertujuan untuk mengevaluasi dampak dari *dropout* dan learning rate terhadap kinerja model BiLSTM-ANN dalam mendeteksi ujaran kebencian di media sosial Indonesia. Penelitian ini berfokus pada dataset dari TikTok dan Twitter (X). Diharapkan, penelitian ini akan memberikan wawasan berharga untuk meningkatkan akurasi model dalam mendeteksi ujaran kebencian di media sosial. Adapun manfaat penelitian ini adalah penelitian ini berkontribusi dalam pengembangan model deep learning untuk klasifikasi teks berbasis BiLSTM-ANN, hasil penelitian ini dapat diterapkan dalam sistem deteksi ujaran kebencian yang lebih efektif untuk meningkatkan moderasi konten di platform media sosial Indonesia dan penelitian ini berpotensi membantu mengurangi penyebaran ujaran kebencian dan dampak negatifnya terhadap masyarakat. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi yang signifikan dalam pengembangan teknologi deteksi ujaran kebencian yang lebih andal, khususnya dalam konteks media sosial di Indonesia.

METODE PENELITIAN

Penelitian ini menggunakan metodologi CRISP-DM untuk mendeteksi ujaran kebencian dalam 2064 komentar dari TikTok dan Twitter (X) sebagai respons terhadap postingan Gibran Rakabuming pada tahun 2024. Komentar tersebut dilabeli sebagai ujaran kebencian atau bukan ujaran kebencian dan diklasifikasikan menggunakan model BiLSTM-ANN yang dioptimalkan dengan menyesuaikan *dropout* dan learning rate. Pengolahan data dan pelatihan model dilakukan menggunakan Python di Google Colab.

Proses penelitian mengikuti metodologi CRISP-DM (Farell, Latt, Jalinus, Yulastri, & Wahyudi, 2024), yang dijelaskan dalam diagram berikut, menggambarkan tahap-tahap utama dalam alur kerja penelitian.



Gambar 1. Alur Penelitian

1. Business Understanding

Pada tahap ini, tujuan utama penelitian didefinisikan, yaitu untuk menilai dampak dari *dropout* dan learning rate terhadap kinerja model BiLSTM-ANN dalam mendeteksi ujaran kebencian di media sosial.

2. Data Understanding

Dataset yang digunakan dalam penelitian ini terdiri dari 2064 komentar yang diambil dari platform TikTok dan Twitter (X), khususnya menanggapi postingan yang dibuat oleh Gibran Rakabuming pada tahun 2024. Pemahaman awal tentang data diperoleh dengan menganalisis distribusi dan karakteristiknya, terutama sifat tekstual data yang berkaitan dengan ujaran kebencian.

3. Data Preparation

Setelah dataset dikumpulkan, dilabeli, dan diproses, langkah-langkah preprocessing dilakukan untuk meningkatkan kinerja model. Pertama, case folding diterapkan, yang mengubah semua teks menjadi huruf kecil untuk memastikan konsistensi, menghilangkan perbedaan akibat sensitivitas huruf besar (Amalia & Sibaroni, 2020). Selanjutnya, pembersihan noise dilakukan menggunakan ekspresi reguler untuk membersihkan teks dengan menghapus simbol, tanda baca, dan karakter yang tidak relevan (Iskandar, Maulana, & Bukhori, 2022).

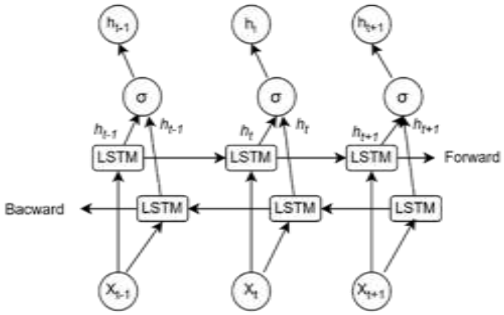
Setelah itu, normalisasi dilakukan untuk menstandarisasi kata-kata yang tidak standar dengan mengubahnya menjadi bentuk yang benar menggunakan kamus kustom (Tahabilder, Islam, & Jahan, 2021). Langkah berikutnya adalah stemming dan penghapusan stopword, di mana kata-kata umum yang tidak memberikan banyak makna dihapus untuk menghindari kebisingan yang tidak perlu dan memungkinkan model untuk fokus pada istilah yang signifikan (Wahyudi & Sibaroni, 2022).

Setelah data dibersihkan dan distandarisasi, embedding kata menggunakan pre-trained FastText dilakukan untuk mengubah kata-kata menjadi vektor numerik yang menangkap makna dan konteks setiap kata (Liu, 2023). Terakhir, tokenisasi dilakukan untuk mengubah teks menjadi urutan token (kata), diikuti dengan padding untuk

memastikan semua urutan memiliki panjang yang sama sebelum dimasukkan ke dalam model BiLSTM-ANN.

4. *Modelling*

Model BiLSTM (Bidirectional Long Short-Term Memory) adalah pengembangan dari LSTM tradisional, yang dirancang untuk memproses data dalam kedua arah, maju dan mundur (Af'idah et al., 2023). Pendekatan dua arah ini memungkinkan model untuk menangkap konteks dari kedua ujung urutan, meningkatkan kemampuannya dalam memahami konteks penuh dari input dan, pada gilirannya, meningkatkan akurasi klasifikasinya.



Gambar 1. Struktur Jaringan Internal BiLSTM

Pada model *Long Short-Term Memory* (LSTM), kondisi tersembunyi pada setiap waktu t disebut h_t Jaringan ini bisa dianggap sebagai LSTM dengan satu lapisan, di mana $\vec{h}_{t-1}$ adalah keadaan tersembunyi pada waktu sebelumnya, yaitu $t-1$. Proses untuk menghitung $\overrightarrow{h_t}$ berdasarkan input $x_t$di waktu t ditunjukkan oleh persamaan berikut:

$$\overrightarrow{h_t} = LSTM\ (x_t, \vec{h}_{t-1}) \qquad (1)$$

Di sini, $\overrightarrow{h_t}$ adalah keadaan tersembunyi pada waktu t, x_t adalah input yang masuk pada waktu t, dan $\vec{h}_{t-1}$ adalah keadaan tersembunyi dari waktu sebelumnya. Untuk jaringan LSTM yang bekerja secara terbalik (*backward*), keadaan tersembunyi pada waktu t, dilambangkan sebagai $\overleftarrow{h_t}$. Hal ini dijelaskan dengan persamaan berikut:

$$\overleftarrow{h_t} = LSTM\ \left(x_t, \overleftarrow{h_{t-1}}\right) (2)$$

Hasil akhir dari jaringan BiLSTM adalah gabungan dari dua kondisi tersembunyi, yaitu $\overrightarrow{h_t}$ dan $\overleftarrow{h_t}$, yang kemudian membentuk satu kondisi tersembunyi lengkap h_t untuk jaringan tersebut (Syaharuddin, Fatmawati, & Suprajitno, 2022).

Dengan menggabungkan output dari kedua arah, BiLSTM secara efektif menangkap informasi kontekstual baik dari arah depan dan belakang. Untuk lebih meningkatkan kinerja model dalam mengklasifikasikan ujaran kebencian, output dari

lapisan BiLSTM diteruskan melalui Jaringan Syaraf Tiruan (ANN). ANN menyempurnakan fitur-fitur yang diekstraksi oleh BiLSTM dan melakukan tugas klasifikasi akhir, membedakan antara ujaran kebencian dan bukan ujaran kebencian.

Dalam penelitian ini, pelatihan model BiLSTM-ANN melibatkan pengaturan dua hyperparameter utama: learning rate dan *dropout* rate. Learning rate mengatur seberapa besar model menyesuaikan bobotnya pada setiap langkah pelatihan. Jika learning rate terlalu tinggi, model dapat melewati solusi optimal; jika terlalu rendah, model akan membutuhkan waktu terlalu lama untuk konvergen. Untuk mengidentifikasi learning rate yang paling efektif, beberapa nilai diuji, termasuk 0.1, 0.0001, 0.00001, 0.002, dan 0.001, berdasarkan penelitian sebelumnya yang menunjukkan kinerja optimal dari nilai-nilai ini dalam model BiLSTM-ANN (Af'idah et al., 2023; Jepkoech, Mugo, Kenduiywo, & Too, 2021; Liu, 2023; Puttarattanamanee et al., 2023; Syaharuddin et al., 2022).

Selain itu, teknik *dropout* diterapkan untuk mencegah *overfitting*. *Dropout* bekerja dengan menonaktifkan secara acak sebagian neuron selama pelatihan, yang membantu model lebih mudah menggeneralisasi pada data yang belum terlihat. Tingkat *dropout* yang diuji adalah 0.02, 0.2, 0.3, 0.4, 0.5, dan 0.7, karena nilai-nilai ini telah terbukti meningkatkan kinerja model dalam tugas klasifikasi teks (Af'idah et al., 2023; Hayaty et al., 2023; Kowsher et al., 2021; Lestandy, 2023; Nugroho et al., 2021).

5. Evaluasi

Kinerja model dievaluasi menggunakan metrik umum seperti akurasi, precision, recall, F1-score, dan AUC (Area Under the Curve). Akurasi mengukur proporsi prediksi yang benar, termasuk baik true positives dan true negatives [20]. Precision menunjukkan proporsi prediksi positif yang benar dari semua prediksi positif. Recall mengukur proporsi positif yang sebenarnya yang berhasil diidentifikasi dengan benar oleh model (Nugroho et al., 2021).

6. *Deployment*

Hasil dari evaluasi digunakan untuk menerapkan model dengan kinerja terbaik untuk deteksi ujaran kebencian secara real-time menggunakan platform Streamlit. Platform interaktif ini memungkinkan pengguna untuk berinteraksi langsung dengan model dengan memasukkan komentar untuk dianalisis. Setelah komentar diajukan, sistem memprosesnya secara real-time dan memberikan umpan balik segera mengenai apakah komentar tersebut dikategorikan sebagai ujaran kebencian atau tidak.

HASIL DAN PEMBAHASAN

Dataset yang terdiri dari 2064 komentar dikumpulkan dari TikTok dan Twitter (X) yang diambil dari postingan Gibran Rakabuming, telah dianalisis menggunakan API Apify dan dilabeli oleh lima penilai independen sebagai ujaran kebencian (1) atau non-ujaran kebencian (0).

Untuk menilai keandalan anotasi, Fleiss' Kappa digunakan untuk mengukur kesepakatan di antara para anotator. Setelah proses pelabelan selesai, serangkaian langkah

prapemrosesan teks diterapkan untuk mempersiapkan data untuk pelatihan model. Dataset kemudian dibagi menjadi set pelatihan, validasi, dan pengujian dengan rasio 7:1:2, seperti yang diuraikan dalam Tabel 1.

Tabel 1. Pembagian Dataset

Rincian Data	**Data *Training* (70%)**		**Data *Validation* (10%)**		**Data *Testing* (20%)**	
Label	0	1	0	1	0	1
Jumlah	909	535	138	68	265	149
Total	1444		206		41	
	2064					

Sumber: Data diolah

Untuk melatih model BiLSTM-ANN, dataset diproses terlebih dahulu menggunakan tokenisasi, padding, dan embeddings kata FastText untuk mengonversi data teks menjadi representasi numerik yang efektif untuk klasifikasi. Teknik pemberian bobot kelas diterapkan untuk mengatasi ketidakseimbangan dataset, memastikan bahwa kelas minoritas (Ujaran Kebencian) mendapatkan bobot yang lebih tinggi selama pelatihan, sehingga meningkatkan sensitivitas model tanpa mengubah distribusi dataset.

Proses pelatihan melibatkan lima skenario eksperimen, yang masing-masing mengevaluasi konfigurasi *dropout* dan learning rate yang berbeda, dengan tujuan untuk mengidentifikasi kinerja model yang optimal. Arsitektur model berhasil menangkap ketergantungan konteks dengan pemrosesan bidirectional, dengan Artificial Neural Network (ANN) yang melakukan tugas klasifikasi. Pelatihan dilakukan di Google Colab, dan kinerja dimonitor melalui berbagai metrik. Tabel 2 menggambarkan konfigurasi model BiLSTM-ANN yang digunakan selama pelatihan.

Tabel 2. Konfigurasi Model

Parameter	**Nilai**
BiLSTM hidden layers	128, 64
ANN hidden layers	128, 64
Epoch	30
Loss Function	*binary_crossentropy*
Batch Size	32
Optimizer	Adam
Dropout	{0.02, 0.2, 0.3, 0.4, 0.5, 0.7}
Learning Rate	{0.01, 0.001, 0.002, 0.0001, 0.00001}

Sumber: Data diolah

Di antara konfigurasi yang diuji, model dengan performa terbaik menggunakan learning rate 0.002 dan *dropout* rate 0.2, mencapai akurasi 68.12%, presisi 53.71%, recall 82.55%, F1-score 65.08%, dan AUC 76.71% pada dataset *testing*. Tabel 3 merangkum hasil klasifikasi.

Tabel 3. Hasil Data *Testing*

Label	Benar	Salah	Total
1 (Ujaran Kebencian)	123	26	149
0 (Non-Ujaran Kebencian)	159	106	265
Total	282	132	414

Sumber: Data diolah

Model ini berhasil mengklasifikasikan 282 dari 414 sampel (68,12%), yang mengindikasikan keefektifannya dalam mendeteksi ujaran kebencian. Namun, terlepas dari daya ingatnya yang kuat, model ini mengalami masalah misklasifikasi, terutama dalam mengidentifikasi komentar-komentar yang non-ujaran kebencian, seperti komentar yang mengandung sarkasme atau kritik tidak langsung.

Kesalahan klasifikasi ini menunjukkan bahwa meskipun model ini efisien dalam mengidentifikasi contoh-contoh ujaran kebencian yang jelas, model ini kesulitan dalam mengidentifikasi bentuk-bentuk ekspresi yang lebih bernuansa atau tidak langsung yang lebih sulit untuk dideteksi. Contoh-contoh sampel yang diklasifikasikan dengan benar dan salah disajikan pada Tabel 4 dan Tabel 5.

Tabel 4. Data Yang Berlabel Benar

No	Text	*Actual*	*Predicted*
1	Ilmu nol adab kosong fufufafa kopong. Bikin malu, mundur aja dari jabatan yang sekarang, ga ada pantas-pantasnya	1	1
2	Mas Gibran model rambutnya belah sekarang	0	0
3	Bisa dilihat beberapa perubahan yang terjadi	0	0

Sumber: Data diolah

Tabel 5. Data Yang Berlabel Salah

No	Teks	Actual	Predicted
1	Cape sekali aku lihat editannya	0	1
2	Gibran Rakabuming presiden	0	1
3	Akhirnya Indonesia punya wakil presiden juga	0	1

Sumber: Data diolah

Lima skenario pelatihan memberikan wawasan yang berharga tentang bagaimana variasi tingkat dropout dan learning rate mempengaruhi kinerja model. Terlihat bahwa meskipun learning rate yang lebih rendah meningkatkan stabilitas model, mencegahnya dari overshooting selama pelatihan, tingkat dropout yang terlalu tinggi memiliki efek yang merugikan pada efektivitasnya.

Tingkat dropout yang lebih tinggi menyebabkan hilangnya informasi berharga secara signifikan selama pelatihan, yang menghambat kemampuan model untuk

menggeneralisasi dengan baik pada data baru. Sebaliknya, hasil penelitian menunjukkan bahwa tingkat dropout moderat sebesar 0,2, ketika dikombinasikan dengan learning rate sebesar 0,002, memberikan keseimbangan optimal antara regularisasi dan efisiensi pembelajaran, yang memungkinkan model untuk konvergen lebih cepat sambil menghindari overfitting. Kombinasi ini ditemukan untuk mencapai kinerja terbaik dalam hal akurasi dan kemampuan generalisasi.

Selain itu, model dengan kinerja terbaik diintegrasikan ke dalam sistem deteksi ujaran kebencian secara real-time, yang dirancang menggunakan template Streamlit. Sistem ini memungkinkan pengguna untuk memasukkan komentar untuk dianalisis dan menerima umpan balik langsung tentang apakah sebuah komentar diklasifikasikan sebagai ujaran kebencian atau tidak. Gambar 3 dan 4 mengilustrasikan antarmuka pengguna sistem, yang menampilkan hasil prediksi bersama dengan skor probabilitas terkait.

Kemampuan sistem untuk mengklasifikasikan komentar secara instan menunjukkan aplikasi praktisnya, menyediakan alat yang efisien untuk pemantauan dan moderasi konten secara *real-time* di platform media sosial. Integrasi model ini ke dalam sistem semacam itu memiliki implikasi penting untuk meningkatkan akurasi dan daya tanggap moderasi konten otomatis, terutama dalam mendeteksi ujaran kebencian di lingkungan online yang beragam.

Gambar 2. Contoh Sistem Mendeteksi Teks Mengandung Ujaran Kebencian

Pada gambar ini, sistem mendeteksi komentar yang mengandung ujaran kebencian dan mengklasifikasikannya dengan benar, dengan memberikan skor probabilitas tinggi sebesar 0,75. Antarmuka dengan jelas menunjukkan komentar yang dimasukkan, dan setelah sistem memprosesnya, hasilnya menunjukkan bahwa komentar tersebut telah diidentifikasi secara akurat sebagai ujaran kebencian.

Probabilitas yang tinggi ini mencerminkan kinerja model yang kuat dalam mengenali konten berbahaya.

Gambar 3. Contoh Sistem Mendeteksi Teks Mengandung Non-Ujaran Kebencian

Gambar ini menunjukkan sistem mendeteksi komentar yang tidak mengandung ujaran kebencian. Sistem dengan tepat mengklasifikasikannya sebagai ujaran kebencian dengan probabilitas rendah (0,11).

Gambar-gambar ini menyoroti kemampuan sistem untuk melakukan klasifikasi ujaran kebencian secara real-time, memberikan umpan balik langsung kepada pengguna apakah sebuah komentar mengandung ujaran kebencian atau tidak. Meskipun model ini bekerja dengan baik, masih ada tantangan dalam menangani komentar yang ambigu dan menangani nuansa halus seperti sarkasme.

Hasil dari penelitian ini menunjukkan bahwa meskipun model BiLSTM-ANN dapat mendeteksi ujaran kebencian dengan recall yang tinggi, model ini mengalami kesulitan dalam mengidentifikasi komentar-komentar yang tidak mengandung ujaran kebencian, seperti komentar-komentar yang mengandung sarkasme atau kritik tidak langsung. Tantangan utama yang dihadapi adalah ketidakseimbangan data dan variasi bahasa yang ditemukan di media sosial. Misalnya, bahasa gaul yang sering digunakan di platform seperti TikTok atau Twitter (X) mengandung nuansa yang sulit ditangkap oleh model.

KESIMPULAN

Penelitian ini berhasil mengimplementasikan model BiLSTM-ANN untuk mendeteksi ujaran kebencian dalam bahasa Indonesia, dengan performa optimal yang dicapai pada *learning rate* 0.002 dan *dropout* rate 0.2. Model ini mencapai akurasi 68,12%, presisi 53,71%, recall 82,55%, F1-score 65,08%, dan AUC 76,71%. Analisis menunjukkan bahwa *learning rate* yang terlalu tinggi (0,01) atau sangat rendah (0,0001, 0,00001) secara signifikan mengganggu kinerja, sementara tingkat *dropout* yang lebih tinggi dari 0,7 mengganggu akurasi klasifikasi. Sebaliknya, tingkat *dropout* yang moderat (0,02 hingga 0,5) menghasilkan kinerja yang lebih baik. Kombinasi learning rate 0,002 dan *dropout* rate 0,2 terbukti paling efektif, mencapai AUC tertinggi dan kehilangan validasi terendah, sehingga menunjukkan kemampuan model yang kuat untuk membedakan ujaran kebencian dan non-kebencian, bahkan dengan set data yang tidak seimbang.

DAFTAR PUSTAKA

abdrakhmanov, Rustam, Kenesbayev, Serik Muktarovich, Berkimbayev, Kamalbek, Toikenov, Gumyrbek, Abdrashova, Elmira, Alchinbayeva, Oichagul, & Ydyrys, Aizhan. (2024). Offensive Language Detection On Social Media Using Machine Learning. *International Journal Of Advanced Computer Science & Applications*, *15*(5).

Af'idah, Dwi Intan, Anggraeni, Puput Dewi, Rizki, Muhammad, Setiawan, Aji Bagus, & Handayani, Sharfina Febbi. (2023). Aspect-Based Sentiment Analysis For Indonesian Tourist Attraction Reviews Using Bidirectional Long Short-Term Memory. *Juita : Jurnal Informatika*, *11*(1), 27. Https://Doi.Org/10.30595/Juita.V11i1.15341

Amalia, Chindy, & Sibaroni, Yuliant. (2020). *Analisis Sentimen Data Tweet Menggunakan Model Jaringan Saraf Tiruan Dengan Pembobotan Delta Tf-Idf*. *7*(2), 7810–7820.

Farell, Geovanne, Latt, Cho Nwe Zin, Jalinus, Nizwardi, Yulastri, Asmar, & Wahyudi, Rido. (2024). Analysis Of Job Recommendations In Vocational Education Using The Intelligent Job Matching Model. *International Journal On Informatics Visualization*, *8*(1), 361–367. Https://Doi.Org/10.62527/Joiv.8.1.2201

Hayaty, Mardhiya, Laksito, Arif Dwi, & Adi, Sumarni. (2023). Hate Speech Detection On Indonesian Text Using Word Embedding Method-Global Vector. *Iaes International Journal Of Artificial Intelligence*, *12*(4), 1928–1937. Https://Doi.Org/10.11591/Ijai.V12.I4.Pp1928-1937

Hesaputra, Akmal Perdana. (2023). Klasifikasi Pelanggaran Undang-Undang Ite Pada Twitter Menggunakan Lstm Dan Bilstm. *Automata, Vol. 4 No.*(Research), 7.

Insani, Zia. (2023). *Analisis Sentimen Komentar Berdasarkan Geo Tagged Menggunakan Algoritma Bilstm*. *10*(1), 211–218.

Iskandar, Thariq, Maulana, Zulkarnain, & Bukhori, Arif Farhan. (2022). *Model Klasifikasi Berbasis Multiclass Classification Dengan Kombinasi Indobert Embedding Dan Long Short- Term Memory Untuk Tweet Berbahasa Indonesia (Classification Model Based On Multiclass Classification With A Combination Of Indobert Embedding And Lon*. *1*(1), 1–28.

Jepkoech, Jennifer, Mugo, David Muchangi, Kenduiywo, Benson K., & Too, Edna Chebet. (2021). The Effect Of Adaptive Learning Rate On The Accuracy Of Neural Networks. *International Journal Of Advanced Computer Science And Applications*, *12*(8), 736–751. Https://Doi.Org/10.14569/Ijacsa.2021.0120885

Kowsher, Md, Tahabilder, Anik, Islam Sanjid, Md Zahidul, Prottasha, Nusrat Jahan, Uddin, Md Shihab, Hossain, Md Arman, & Kader Jilani, Md Abdul. (2021). Lstm-Ann & Bilstm-Ann: Hybrid Deep Learning Models For Enhanced Classification Accuracy. *Procedia Computer Science*, *193*, 131–140.

Https://Doi.Org/10.1016/J.Procs.2021.10.013

Lestandy, Merinda. (2023). Exploring The Impact Of Word Embedding Dimensions On Exploring The Impact Of Word Embedding Dimensions On Depression Data Classification Using Bilstm Model Depression Data Classification Using Bilstm Model. *Procedia Computer Science*, *227*, 298–306. Https://Doi.Org/10.1016/J.Procs.2023.10.528

Liu, Jinjun. (2023). Sentiment Classification Of Social Network Text Based On At-Bilstm Model In A Big Data Environment. *International Journal Of Information Technologies And Systems Approach*, *16*(2), 1–15. Https://Doi.Org/10.4018/Ijitsa.324808

Nugroho, Kuncahyo Setyo, Akbar, Ismail, Suksmawati, Affi Nizar, Komputer, Ilmu, Brawijaya, Universitas, Mada, Universitas Gadjah, Teknik, Fakultas, & Malang, Universitas Widyagama. (2021). *Deteksi Depresi Dan Kecemasan Pengguna Twitter Menggunakan Bidirectional Lstm*. (Ciastech), 287–296.

Puttarattanamanee, Maneerat, Boongasame, Laor, & Thammarak, Karanrat. (2023). A Comparative Study Of Sentiment Analysis Methods For Detecting Fake Reviews In E-Commerce. *Hightech And Innovation Journal*, *4*(2), 349–363. Https://Doi.Org/10.28991/Hij-2023-04-02-08

Setiawan, Esther Irawati, & Lestari, Ika. (2021). Stance Classification Pada Berita Berbahasa Indonesia Berbasis Bidirectional Lstm. *Journal Of Intelligent System And Computation*, *3*(1), 41–48. Https://Doi.Org/10.52985/Insyst.V3i1.148

Syaharuddin, Syaharuddin, Fatmawati, Fatmawati, & Suprajitno, Herry. (2022). Best Architecture Recommendations Of Ann Backpropagation Based On Combination Of Learning Rate, Momentum, And Number Of Hidden Layers. *Jtam (Jurnal Teori Dan Aplikasi Matematika)*, *6*(3), 629. Https://Doi.Org/10.31764/Jtam.V6i3.8524

Tahabilder, Anik, Islam, Zahidul, & Jahan, Nusrat. (2021). Sciencedirect Sciencedirect Lstm-Ann & Bilstm-Ann : Hybrid Deep Learning Models For Enhanced Classification Accuracy. *Procedia Computer Science*, *193*, 131–140. Https://Doi.Org/10.1016/J.Procs.2021.10.013

Toktarova, Aigerim, Syrlybay, Dariga, Myrzakhmetova, Bayan, Anuarbekova, Gulzat, Rakhimbayeva, Gulbarshin, Zhylanbaeva, Balkiya, Suieuova, Nabat, & Kerimbekov, Mukhtar. (2023). Hate Speech Detection In Social Networks Using Machine Learning And Deep Learning Methods. *International Journal Of Advanced Computer Science And Applications*, *14*(5), 396–406. Https://Doi.Org/10.14569/Ijacsa.2023.0140542

Wahyudi, Diki, & Sibaroni, Yuliant. (2022). Deep Learning For Multi-Aspect Sentiment Analysis Of Tiktok App Using The RNN-LSTM Method. *Building Of Informatics, Technology And Science (BITS)*, *4*(1), 169–177. Https://Doi.Org/10.47065/Bits.V4i1.1665