Harap mengisi tabel ini, Tabel ini digunakan untuk keperluan komunikasi administrasi saja, saat publish akan dihapus oleh team editor.	
Nama Kontak	
Nomor WA	
Prodi/Jurusan	
Perguruan Tinggi	

Klasifikasi Ujaran Kebencian di Media Sosial Berbasis Metode Long Short-Term Memory (LSTM)

Abiyanfaras Damuyasa[1], Birrham Efendi Lubis[2], Hafidza Dafariz Mujizat[3], Muhammad Zidan Fadillah[4]
Teknik Informatika, Institut Teknologi Nasional Malang
Jalan Raya Karanglo km 2 Malang, Indonesia
birrhamefendilubis@mhs.pelitabangsa.ac.id

ABSTRAK

Cepatnya pemanfaatan media sosial, terutama Twitter, telah mendorong penyebaran ujaran kebencian (hate speech), yang bisa memicu konflik sosial yang meluas. Secara teknis, identifikasi ujaran kebencian merupakan tugas Klasifikasi Teks Biner dalam NLP, di mana pendekatan konvensional sering kali tidak dapat mengerti konteks semantik yang rumit. Masalah utama yang dihadapi adalah ketidakseimbangan kelas dalam dataset, di mana kelas Non-Hate jauh lebih mendominasi (93,0% dibandingkan 7,0%). Studi ini bertujuan untuk menciptakan model klasifikasi ujaran kebencian yang tahan banting dengan menyelesaikan isu ketidakseimbangan data. Metode yang diterapkan adalah arsitektur Bidirectional Long Short-Term Memory (Bi-LSTM), yang dirancang agar dapat menangkap konteks sekuensial dari dua arah. Strategi Oversampling diterapkan pada Set Pelatihan (80% data) untuk mendapatkan distribusi kelas yang seimbang (50% lawan 50%). Hasil uji coba pada Validation Set menunjukkan model memperoleh Akurasi Weighted Average sebesar 0.96. Walaupun kinerja secara keseluruhan sangat memuaskan, analisis metrik Kelas 1 (Hate/Offensive) menunjukkan nilai Recall yang rendah (0.58), yang disebabkan oleh banyaknya False Negative (186 kasus), menunjukkan bahwa sensitivitas model masih dipengaruhi oleh ketidakseimbangan dalam data validasi asli
Kata kunci : *tuliskan maksimum 6 kata kunci di sini*

1. PENDAHULUAN

Perkembangan cepat teknologi informasi, terutama melalui platform media sosial seperti Twitter, telah secara mendasar mengubah wajah komunikasi global. Platform ini menyediakan sarana untuk bertukar informasi secara langsung dan menghasilkan ruang publik digital untuk miliaran pengguna. Akan tetapi, kemudahan untuk anonim dan penyebaran informasi yang cepat juga memiliki dampak buruk, salah satunya adalah peningkatan prevalensi ujaran kebencian (hate speech)[1]. Ujaran kebencian diartikan sebagai ungkapan yang menyerang orang atau kelompok berdasarkan karakteristik tertentu seperti ras, agama, etnis, atau jenis kelamin, dan memiliki potensi untuk memicu konflik sosial yang luas serta cepat menyebar [1].

Secara teknis, pengenalan ujaran kebencian adalah tugas klasik dalam Pemrosesan Bahasa Alami (NLP), di mana sebuah teks dikelompokkan menjadi kategori ujaran kebencian (kelas 1) atau non-ujaran kebencian (kelas 0). Beragam studi telah dilakukan untuk menyelesaikan isu ini, mulai dari metode Machine Learning (ML) konvensional seperti Naïve Bayes dan N-Gram, sampai penerapan Contextual Embedding [2]. Walaupun metode ML konvensional menunjukkan hasil yang cukup baik, seringkali mereka mengalami kesulitan dalam menangkap fitur semantik yang rumit, urutan kata, dan ketergantungan jarak jauh yang penting dalam konteks bahasa yang kompleks dan dinamis seperti di platform media social.

Untuk mengatasi kekurangan itu dan mencari solusi yang lebih kuat, penelitian ini menekankan pada pemanfaatan arsitektur Long Short-Term Memory (LSTM). Model Bidirectional-LSTM (Bi-LSTM) telah terbukti efektif dalam mengolah data sekuensial dan berhasil mendeteksi berbagai jenis pelanggaran digital, termasuk ujaran kebencian, berkat kemampuannya memahami konteks dari dua arah [3].

2. TINJAUAN PUSTAKA

Tinjauan pustaka ini mengulas konsep-konsep dasar yang berkaitan dengan pengelompokan ujaran kebencian di platform media sosial, evaluasi terhadap metode pengelompokan yang telah ada, serta menyoroti peran jaringan saraf tiruan deep learning, terutama arsitektur Long Short-Term Memory (LSTM)

2.1. Konsep Ujaran Kebencian (Hate Speech) dan Klasifikasi Teks

Ujaran Kebencian (*Hate Speech*) didefinisikan secara umum sebagai ungkapan lisan atau tulisan yang bertujuan untuk mengejek, menghina, atau merendahkan orang atau kelompok tertentu berdasarkan karakteristik yang tidak dapat diubah

seperti ras, agama, etnis, atau orientasi seksual.[1], [2]. Dalam konteks Pemrosesan Bahasa Alami (NLP), pengenalan ujaran kebencian merupakan tugas Klasifikasi Teks Biner, di mana teks yang diberikan (tweet) harus dikategorikan ke dalam salah satu dari dua kelas: Ujaran Kebencian (Kelas 1) atau Tidak Ujaran Kebencian (Kelas 0).

2.2. Pra-pemrosesan Teks

Kualitas teks yang dimasukkan sangat berpengaruh terhadap kinerja model. Dalam klasifikasi ujaran kebencian, langkah Pra-pemrosesan Teks sangat penting karena data dari media sosial mengandung banyak kebisingan seperti tautan URL, sebutan (*@user*), tanda baca, dan penggunaan bahasa non-standar (slang) [2]. Proses pra-pemrosesan yang umum meliputi:

1. Pembersihan Noise : Penghapusan URL, penyebutan, angka, dan tanda baca.
2. Normalisasi : Mengubah seluruh teks menjadi huruf kecil (penurunan huruf besar).
3. Stopword Removal : Menghapus kata-kata yang tidak memberikan makna semantik yang berarti.
4. Lemmatisasi : Mengonversi kata ke bentuk dasar untuk mengurangi ukuran kosakata

2.3. Metode Klasifikasi Teks dalam Deteksi Ujaran Kebencian

Studi sebelumnya tentang deteksi ujaran kebencian telah menerapkan berbagai metode :

2.3.1. Metode Tradisional dan Keterbatasannya

Metode Machine Learning (ML) konvensional, seperti Naïve Bayes yang dipadukan dengan ekstraksi fitur N-Gram, biasanya digunakan sebagai acuan karena kesederhanaan dan efisiensi komputasinya [1]. Akan tetapi, ML konvensional cenderung memperhatikan kurang pada urutan kata dan struktur kalimat, sehingga mengalami kesulitan dalam menangkap konteks semantik yang sangat penting, terutama pada teks yang kompleks dan ambigu seperti ujaran kebencian di Twitter. Hal ini mengarahkan pergeseran menuju model yang dapat memahami konteks yang lebih mendalam, seperti pendekatan Embedding Kontekstual [2].

2.3.2. Metode Deep Learning (*LSTM)*

Long Short-Term Memory (LSTM) merupakan arsitektur Recurrent Neural Network (RNN) yang hebat karena mekanisme gating-nya, sehingga dapat mengatasi kelemahan RNN konvensional dalam menangani ketergantungan jangka panjang. Pada klasifikasi teks, Bidirectional-LSTM (Bi-LSTM) terbukti mampu menghasilkan representasi kontekstual dua arah yang kaya, yang sangat vital untuk mengidentifikasi ujaran kebencian dan pelanggaran UU ITE [3]. Kemampuan LSTM ini tidak hanya terbatas pada ujaran kebencian tetapi juga terbukti efektif dalam pengklasifikasian konten berbahaya lainnya seperti Cyberbullying di platform media social [4]. Beragam penelitian menguatkan kuatnya performa LSTM dalam mendeteksi ujaran kebencian di Twitter [5], bahkan melalui optimasi menggunakan Algoritma Genetika (GA) [6], menegaskan model ini sebagai fondasi kuat untuk tugas ini.

2.4. *Research Gap*

Walaupun model LSTM menunjukkan performa yang unggul, salah satu tantangan yang selalu ada dalam literatur deteksi ujaran kebencian adalah ketidakseimbangan kelas pada dataset [1]. Mayoritas studi lebih menekankan pada desain model, namun sering kali melupakan cara yang tepat untuk menangani ketidakseimbangan. Penelitian ini bertujuan untuk mengatasi kesenjangan ini dengan secara jelas mengintegrasikan metode LSTM bersama teknik Oversampling pada data pelatihan. Tujuannya adalah untuk menghasilkan distribusi kelas yang seimbang, menjamin model tidak hanya akurat secara keseluruhan, tetapi juga efisien dan tanpa bias dalam mengenali kelas minoritas (ujaran kebencian).

3. METODE PENELITIAN

Bagian ini menjelaskan langkah-langkah penelitian yang dilakukan, mulai dari pengumpulan dan pemrosesan awal data, strategi menangani ketidakseimbangan kelas, sampai perancangan model Long Short-Term Memory (LSTM) serta metrik evaluasi yang dipakai

3.1 Sumber Data dan Pra-Pemrosesan Awal

3.1.1. Sumber Data

Studi ini memanfaatkan dataset publik yang berbahasa Indonesia yang diambil dari platform Twitter. Dataset ini dapat diunduh melalui platform Kaggle [7], yang biasa digunakan sebagai tempat penyimpanan data ujaran kebencian dan bahasa yang tidak sopan. Dataset dibagi menjadi dua kelas biner: Non-Hate/Offensive (0) dan Hate/Offensive (1)

3.1.2. Pra-Pemrosesan Teks

Dalam rangka menghilangkan noise khas yang ada di media sosial dan menstandarisasi teks, sebuah urutan langkah pra-pemrosesan diterapkan [1]. Proses ini mencakup: Penghapusan Noise

(menghilangkan URL, sebutan, dan angka), Penyeragaman Huruf Kecil, Penghapusan Stopword (menghapus kata-kata umum), dan Lemmatisasi (kembali ke bentuk dasar kata)

3.2 Strategi Penanganan Ketidakseimbangan Data (Oversampling)

Dataset ujaran kebencian secara alami menghadapi ketidakseimbangan kelas. Untuk menghindari bias pada model, teknik Oversampling diterapkan hanya pada Training Set. Data dibagi menjadi 80% untuk Kumpulan Pelatihan dan 20% untuk Kumpulan Validasi (80:20) menggunakan sampling terestruktur. Kelas minoritas (Hate/Offensive) di-oversample dengan melakukan sampling dengan penggantian sampai jumlahnya sama dengan kelas mayoritas [8]. Strategi ini penting untuk memperkuat kemampuan model dalam mengenali ujaran kebencian

3.3 Tokenisasi dan *Embedding*

Teks yang sudah dibersihkan disusun ke dalam format yang bisa diproses oleh model :

- **Tokenisasi dan Padding**: Memanfaatkan Tokenizer (Keras) dengan limitasi maksimum 10.000 kata unik (max_words). Panjang urutan token disamakan menjadi 100 token (max_len) dengan menggunakan Padding [9].
- **Representasi Target (Label)**: Label biner (0 dan 1) diubah menjadi format One-Hot Encoding (OHE) agar sesuai dengan fungsi *loss categorical_crossentropy.*

3.4 Arsitektur Model Deep Learning (Bi-LSTM)

Model yang diterapkan adalah arsitektur Sequential yang menggunakan *Bidirectional Long Short-Term Memory* (Bi-LSTM), dirancang untuk menangkap konteks sekuensial teks dari dua arah (ke depan dan ke belakang). Model dimulai dengan Lapisan *Embedding* yang mengubah token menjadi vektor padat berdimensi **32**. Lapisan ini terhubung dengan lapisan inti, yaitu **Bi-LSTM** yang memiliki **16** ***unit***, yang berfungsi untuk memproses ketergantungan jangka panjang dalam teks. Hasil dari Bi-LSTM selanjutnya diberikan kepada Lapisan ***Dense*** yang memiliki ***512 unit*** dan fungsi aktivasi ReLU, dilengkapi dengan *regularizer kernel L1* untuk menghindari *overfitting*. Arsitektur dilanjutkan dengan lapisan Normalisasi *Batch* dan *Dropout* dengan tingkat **0.3** untuk meningkatkan kestabilan serta generalisasi model. Lapisan terakhir merupakan Lapisan *Dense Output* yang memiliki *2 unit* dengan aktivasi *Softmax*, menghasilkan probabilitas akhir untuk klasifikasi biner (Hate/Non-Hate). Model ini dikompilasi dengan fungsi loss *categorical_crossentropy* dan *optimizer adam.*

3.5 Pelatihan dan Evaluasi Model

Model dilatih selama 50 epoch dengan ukuran batch 32. Teknik Early Stopping (ES) dan ReduceLROnPlateau (LR) diterapkan untuk mengoptimalkan dan menghentikan pelatihan apabila akurasi validasi tidak mengalami kenaikan selama 3 epochs dan jika kerugian validasi tidak berubah selama 2 epochs. Evaluasi dilakukan pada Validation Set dengan menggunakan metrik menyeluruh seperti Akurasi, Precision, Recall, dan F1-Score [6], , yang krusial untuk mengevaluasi kinerja pada data yang tidak seimbang,

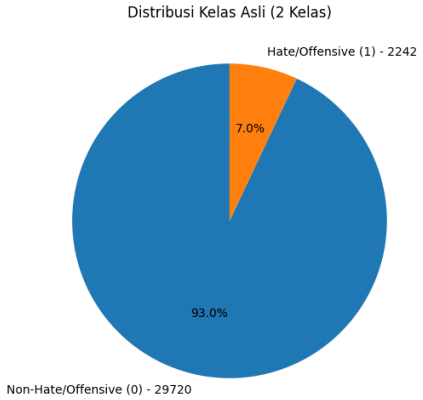
4. HASIL DAN PEMBAHASAN

Bagian ini menunjukkan hasil dari penerapan model Bidirectional Long Short-Term Memory (Bi-LSTM) yang sudah dilatih pada data Twitter seimbang menggunakan teknik Oversampling. Pembahasan tertuju pada analisis penyebaran data, kinerja model di data validasi, serta penafsiran metrik evaluasi.

4.1 Analisis Data dan Pra-Pemrosesan

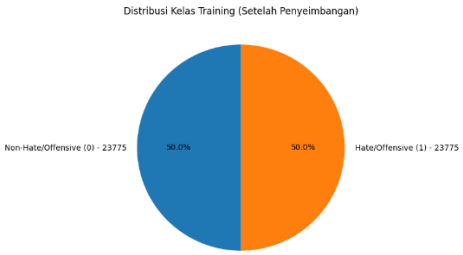
4.1.1 Distribusi Kelas

Analisis awal terhadap dataset mengindikasikan adanya masalah ketidakseimbangan kelas yang cukup signifikan. Dari jumlah total sampel data, Kelas 0 (Non-Hate/Offensive) mendominasi dengan proporsi 93,0% (29.720 sampel), sedangkan Kelas 1 (Hate/Offensive) hanya mencakup 7,0% (2.242 sampel). Perbandingan visual ini diperlihatkan di Gambar 4.1.



Gambar 4.1 Distribusi Kelas Dataset Asli (Tidak Seimbang)

Untuk mengurangi bias model, teknik Oversampling digunakan pada Training Set (80% data). Setelah penyeimbangan, Training Set berhasil memperoleh distribusi yang ideal, yaitu 50.0% untuk setiap kelas, dengan jumlah total 23.775 sampel per kelas. Distribusi kelas yang seimbang pada data latih ini divisualisasikan dalam Gambar 4.2



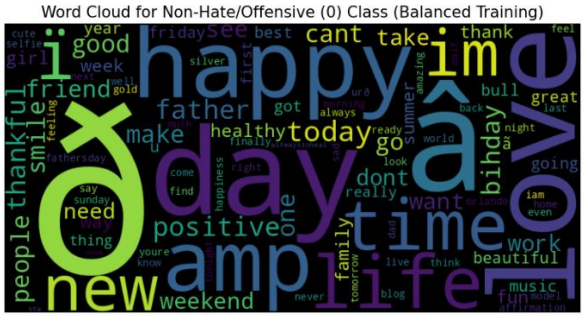
Gambar 4.2 Distribusi Kelas pada Training Set (Setelah Oversampling)

4.1.2 Karakteristik Fitur Teks

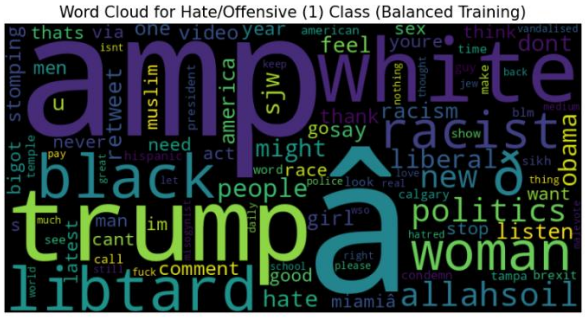
Analisis Word Cloud memberikan gambaran tentang kata-kata yang paling umum muncul setelah proses pra-pemrosesan, membedakan pola bahasa di antara kedua kategori:

- Kelas Non-Hate/Offensive (0): Kata-kata yang lebih banyak menunjukkan sentimen positif atau netral, seperti 'bahagia', 'hari', dan 'baik'.
- Kelas Kebencian/Serangan (1): Kata-kata yang mencolok adalah istilah politik, bahasa kasar, atau ungkapan yang mengandung perasaan negatif yang mendalam.

Visualisasi Word Cloud untuk kedua kategori ditampilkan pada Gambar 4.3 dan Gambar 4.4, menegaskan keberhasilan proses Embedding dalam menangkap fitur bahasa yang penting.



Gambar 4.3 *Word Cloud* untuk Kelas *Non-Hate/Offensive* (0)



Gambar 4.4 *Word Cloud* untuk Kelas Hate/Offensive (1)

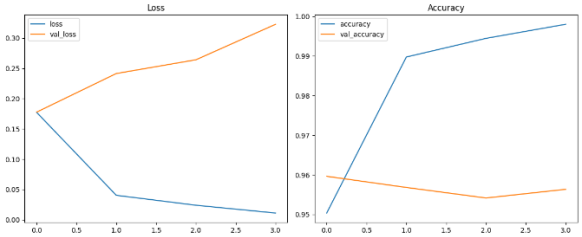
4.2 Kinerja Model dan Analisis Pelatihan

4.2.1 Curva *Loss* dan *Accuracy*

Model Bi-LSTM dilatih selama 50 epoch, namun pelatihan dihentikan lebih awal (Early Stopping) karena performa pada data validasi tidak menunjukkan peningkatan yang signifikan.

Grafik di Gambar 4.5 memperlihatkan perkembangan Loss dan Accuracy selama tahap pelatihan:

- Loss: Kerugian pelatihan (loss) menurun drastis, mencapai nilai yang sangat rendah, yang menunjukkan model telah beradaptasi dengan baik terhadap data latih yang seimbang. Namun, terdapat perbedaan antara training loss dan validation loss (val_loss), di mana val_loss menunjukkan tren kenaikan setelah epoch pertama, mengindikasikan adanya sedikit overfitting pada data validasi yang masih tidak seimbang
- Acuracy: Akurasi pelatihan mencapai nilai yang sangat tinggi (sekitar 99.6%),sedangkan akurasi validasi (val_accuracy) menunjukkan nilai yang stabil di sekitar 95.6%



Gambar 4.5 Kurva *Loss* dan *Accuracy* Selama Pelatihan

4.2.2 Evaluasi Kinerja (Classification Report)

Model inerja diukur lewat Classification Report pada Validation Set (total 6.393 sampel) yang hasilnya ditampilkan pada Gambar 4.6. Secara keseluruhan, model memperoleh Akurasi (Rata-rata Berbobot) sebesar 0.96

```
--- Classification Report ---
                    precision    recall  f1-score   support

Non-Hate/Offensive (0)       0.97      0.99      0.98      5945
    Hate/Offensive (1)       0.78      0.58      0.67       448

              accuracy                           0.96      6393
             macro avg       0.88      0.79      0.82      6393
          weighted avg       0.96      0.96      0.96      6393
```

Gambar 4.6 *Classification Report* Model Bi-LSTM

Meskipun akurasi keseluruhan tinggi, metrik untuk Kelas 1 (Hate/Offensive) menunjukkan tantangan:

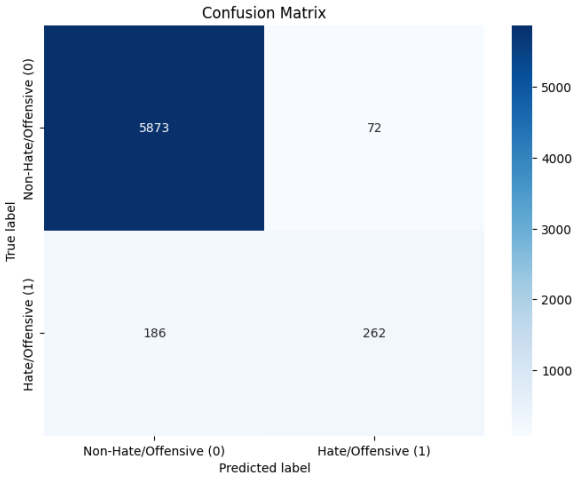
- Akurasi (0.78): Tingkat akurasi yang cukup baik (78%), menunjukkan bahwa model memiliki kehandalan yang wajar saat

mengklasifikasikan sebuah tweet sebagai ujaran kebencian

- Recall (0.58): Nilai Recall yang cukup rendah ini mengindikasikan bahwa model hanya mampu mendeteksi $\mathbf{58\%}$ dari total ujaran kebencian yang seharusnya terdapat dalam data validasi. Kegagalan model dalam mendeteksi $\mathbf{42\%}$ sampel ujaran kebencian menunjukkan bahwa masalah ketidakseimbangan kelas pada data validasi masih berpengaruh terhadap sensitivitas model

4.3 Analisis *Confusion Matrix*

Hasil dari Confusion Matrix (Gambar 4.7) mengklarifikasi kinerja yang diperoleh pada Classification Report.



Gambar 4.7 *Confusion Matrix* Hasil Klasifikasi

Dari total 6.393 sampel data validasi, didapatkan rincian berikut:

- True Negative (TN) = 5873: Jumlah sampel non-hate yang diprediksi benar sebagai non-hate.
- True Positive (TP) = 262: Jumlah sampel hate yang diprediksi benar sebagai hate.
- False Positive (FP) = 72: Jumlah sampel non-hate yang salah diprediksi sebagai hate (Type I Error).
- False Negative (FN) = 186: Jumlah sampel hate yang salah diprediksi sebagai non-hate (Type II Error).

Tingginya angka False Negative (FN = 186), yang mencakup 41.5% dari total sampel Kelas 1, adalah penyebab utama rendahnya Recall Kelas 1. Ini menunjukkan bahwa meskipun model memiliki kemampuan baik dalam membedakan kelas mayoritas, masih ada keterbatasan dalam sensitivitas untuk mendeteksi seluruh kasus ujaran kebencian, suatu temuan yang sering terjadi dalam penelitian dengan data yang sangat tidak seimbang [10].

5. KESIMPULAN DAN SARAN

5.1. Kesimpulan

Penelitian ini berhasil menciptakan model klasifikasi ujaran kebencian biner pada data Twitter berbahasa Inggris dengan menggunakan metode Bidirectional Long Short-Term Memory (Bi-LSTM) yang dipadukan dengan strategi Oversampling untuk menangani ketidakseimbangan data yang signifikan

1. Efektivitas Penyeimbangan Data: Penerapan Oversampling berhasil menghasilkan Training Set yang seimbang (50% Hate Speech dan 50% Non-Hate), sehingga model Bi-LSTM dapat belajar dengan baik dari kelas yang kurang.
2. Kinerja Model Secara Keseluruhan: Model Bi-LSTM menunjukkan performa yang sangat baik, meraih Akurasi Weighted Average sebesar 0.96 di Validation Set.
3. Tantangan Deteksi Kelas Minoritas: Meskipun akurasi totalnya tinggi, model tetap mengalami kesulitan dalam mendeteksi seluruh kasus ujaran kebencian (Kelas 1). Hal ini terlihat dari nilai Recall yang cukup rendah (0.58) dan nilai F1-Score (0.67) untuk Kelas 1. Analisis Matriks Kebingungan menunjukkan tingginya jumlah False Negative (186 kasus), yang mengindikasikan bahwa model masih cenderung berhati-hati dalam memprediksi kelas minoritas akibat warisan ketidakseimbangan data pada Validation Set yang asli.

Secara ringkas, Long Short-Term Memory (LSTM) dan variannya adalah dasar yang kokoh untuk klasifikasi teks sekuensial, dan dengan penanganan pra-pemrosesan yang tepat, model ini dapat berfungsi sebagai sistem deteksi dini yang efektif

5.2. Saran

Berdasarkan temuan dan keterbatasan yang ada, rekomendasi untuk penelitian selanjutnya mencakup:

1. Eksplorasi Metode Penanganan Ketidakseimbangan yang Lebih Lanjut: Sebagai alternatif dari oversampling dasar, pertimbangkan untuk menggunakan teknik yang lebih kompleks seperti SMOTE (Synthetic Minority Over-sampling Technique) atau pengaturan bobot kelas (class weighting) pada fungsi kerugian untuk memberikan penalti lebih besar terhadap

kesalahan False Negative (FN), yang dapat meningkatkan Recall dari model.

2. Penggunaan Model Transformer: Disarankan untuk melakukan perbandingan uji coba dengan model berbasis Transformer seperti BERT atau IndoBERT, karena model tersebut telah terbukti sangat efisien dalam menangkap konteks linguistik yang rumit dan dapat mengalahkan model RNN seperti Bi-LSTM dalam tugas klasifikasi ujaran kebencian.
3. Penyetelan Hyperparameter: Lakukan penyetelan hyperparameter yang lebih komprehensif, termasuk jumlah unit LSTM, tingkat Dropout, dan Learning Rate model, mungkin dengan bantuan algoritma optimasi seperti Genetic Algorithm (GA), untuk mengoptimalkan kinerja dan kestabilan model.
4. Pengayaan Fitur Leksikal: Tingkatkan tahapan pra-pemrosesan data dengan mengintegrasikan normalisasi bahasa non-formal (slang) atau Pendekatan Penyematan Kontekstual untuk memperkuat representasi makna, terutama pada data media sosial yang sangat bising.

DAFTAR PUSTAKA

[1] I. And and D. Expert, "Deteksi Ujaran Kebencian dengan Metode Klasifikasi Naïve Bayes dan Metode N-Gram pada Dataset Multi-Label Twitter Berbahasa Indonesia INFORMASI ARTIKEL ABSTRAK," 2022. [Online]. Available: http://index.unper.ac.id

[2] G. B. Herwanto, A. M. Ningtyas, I. G. Mujiyatna, K. E. Nugraha, and I. N. Prayana Trisna, "Hate Speech Detection in Indonesian Twitter using Contextual Embedding Approach," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 15, no. 2, p. 177, Apr. 2021, doi: 10.22146/ijccs.64916.

[3] M. Dhafa Maulana, C. Sri, and K. Aditya, "Perbandingan IndoBERT dan Bi-LSTM Dalam Mendeteksi Pelanggaran Undang-Undang ITE", [Online]. Available: https://doi.org/10.31598

[4] K. Chuluq and S. R. Nudin, "Klasifikasi Cyberbullying Pada Media Sosial Dengan Menggunakan Metode Recurrent Neural Network Dan Long Short Term Memory," *Journal of Informatics and Computer Science*, vol. 06, 2024.

[5] R. Y. Rafael and F. Adikara, "PENGIMPLMENTASIAN ALGORITMA LONG SHORT-TERM MEMORY UNTUK MENDETEKSI UJARAN KEBENCIAN PADA APLIKASI TWITTER," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 8, no. 2, pp. 551–560, May 2023, doi: 10.29100/jipi.v8i2.3490.

[6] D. Alfatihah, N. Erlani, and E. B. Setiawan, "Hate Comment Detection On Twitter Using Long Short Term Memory (LSTM) With Genetic Algorithm (GA)," *Journal of Universal Studies*, vol. 4, no. 11, 2024, [Online]. Available: http://eduvest.greenvest.co.id

[7] R. C. Maringka, R. Justino, N. Makarawung, K. Kunci, : Bert, and U. Kebencian, "OPTIMALISASI ANALISIS UJARAN KEBENCIAN ULASAN E-COMMERCE BERBASIS BERT DAN FAISS," 2025.

[8] I. Bagus Jatiarso and Y. Azhar, "Klasifikasi Emosi Pada Tweet Pengguna Platform X Menggunakan Metode LSTM-GloVe Berbasis SMOTE," *REPOSITOR*, vol. 7, no. 3, pp. 401–406, 2025.

[9] E. Aurora Az Zahra, Y. Sibaroni, and S. Suryani Prasetyowati, "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, pp. 170–178, Jul. 2023, doi: 10.35877/454ri.jinav1864.

[10] K. Publik, Y. Termasuk, J. Pelanggaran, N. Qur'atul 'ain, B. Pramono, and A. H. Wibowo, "Penerapan Metode LSTM Pada Sistem Klasifikasi," *ANIMATOR*, vol. 2, no. 1, pp. 1–5.