Jurnal Surya Informatika
Journal homepage: https://jurnal.umpp.ac.id/index.php/surya_informatika
P-ISSN: 2477-3042; E-ISSN: 3026-3034
Vol. 16, No. 2, November 2026, Pp. 267-276

Perbandingan Performa Algoritma Deep Learning dan Machine Learning untuk Klasifikasi Text Aduan di Sosial Media

Muhamad Nasim[1)], Fadilah Safitri[2)], Enggal Meureksa Ibrahim[3*)], Arya Surya Pratama[4)]
1. Program Studi Informatika, Fakultas Saintek, UIN Prof. K.H. Saifuddin Zuhri Purwokerto, Indonesia
2. Program Studi Informatika, Fakultas Saintek, UIN Prof. K.H. Saifuddin Zuhri Purwokerto, Indonesia
3. Program Studi Informatika, Fakultas Saintek, UIN Prof. K.H. Saifuddin Zuhri Purwokerto, Indonesia
4. Program Studi Informatika, Fakultas Saintek, UIN Prof. K.H. Saifuddin Zuhri Purwokerto, Indonesia

Article Info

Kata Kunci: *Aduan; Deep Learning; Komplain; Machine Learning; Natural Language Processing*

***Keywords*:** *Klasifikasi Teks; Aduan Publik; Naive Bayes; Deep Learning; Media Sosial; Natural Language Processing*

Article history:
Received 21 Juli 2026
Revised 9 September 2026
Accepted 19 September 2026
Available online 1 November 2026

DOI :
10.48144/suryainformatika.v16i2.2509

* Corresponding author.
Enggal Meureksa Ibrahim
E-mail address:
244110601018@mhs.uinsaizu.ac.id

ABSTRAK

Perkembangan media sosial menghasilkan volume data aduan publik berbahasa Indonesia yang besar dan beragam sehingga diperlukan metode klasifikasi otomatis yang efektif dan efisien. Penelitian ini bertujuan membandingkan performa algoritma machine learning, yaitu Naive Bayes dan Support Vector Machine (SVM), dengan algoritma deep learning, yaitu Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), dan Generative Adversarial Network (GAN), dalam melakukan klasifikasi teks aduan publik. Kebaruan penelitian ini terletak pada evaluasi penggunaan GAN sebagai model klasifikasi melalui komponen Discriminator pada konteks aduan publik berbahasa Indonesia. Dataset yang digunakan terdiri atas 5.000 data teks yang diperoleh dari media sosial, dengan distribusi 2.127 data Complaint (42,5%) dan 2.873 data Not Complaint (57,5%). Tahapan penelitian meliputi preprocessing teks, pembagian data menggunakan stratified sampling dengan proporsi 80% data latih dan 20% data uji, serta evaluasi model menggunakan Accuracy, Precision, Recall, dan F1-Score berdasarkan rata-rata tiga kali pengujian. Hasil penelitian menunjukkan bahwa Naive Bayes menghasilkan performa terbaik dengan Accuracy sebesar 77,50%, Precision 77,41%, Recall 77,50%, dan F1-Score 77,31%. Model berikutnya adalah CNN dengan F1-Score 77,09%, SVM sebesar 74,67%, LSTM sebesar 73,66%, dan GAN sebesar 41,98%. Rendahnya performa GAN dipengaruhi oleh ketidakstabilan proses pelatihan adversarial, sedangkan LSTM mengalami kecenderungan overfitting pada dataset berukuran sedang. Penelitian ini menunjukkan bahwa Naive Bayes lebih efektif untuk klasifikasi aduan publik dengan jumlah data terbatas karena memiliki akurasi kompetitif dan kompleksitas komputasi yang lebih rendah. Penelitian selanjutnya dapat mengembangkan dataset yang lebih besar, memperluas kategori aduan, menerapkan teknik augmentasi data, serta mengeksplorasi model berbasis transformer seperti BERT dan IndoBERT.

ABSTRACT

The growth of social media has generated a large and diverse volume of Indonesian-language public complaint data, requiring effective and efficient automatic classification methods. This study aims to compare the performance of machine learning algorithms, namely Naive Bayes and Support Vector Machine (SVM), with deep learning algorithms, including Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Generative Adversarial Network (GAN), for public complaint text classification. The novelty of this research lies in evaluating GAN as a classification model through its Discriminator component in the context of Indonesian public complaints. The dataset consists of 5,000 text samples collected from social media platforms, comprising 2,127 Complaint data (42.5%) and 2,873 Not Complaint data (57.5%). The research process includes text preprocessing, data splitting using stratified sampling with an 80% training and 20% testing ratio, and model evaluation using Accuracy, Precision, Recall, and F1-Score based on the average results of three experimental runs. The results indicate that Naive Bayes achieved the best performance with an Accuracy of 77.50%, Precision of

Jurnal Surya Informatika
Journal homepage: https://jurnal.umpp.ac.id/index.php/surya_informatika
P-ISSN: 2477-3042; E-ISSN: 3026-3034
Vol. 16, No. 2, November 2026, Pp. 267-276

77.41%, Recall of 77.50%, and F1-Score of 77.31%. It was followed by CNN with an F1-Score of 77.09%, SVM with 74.67%, LSTM with 73.66%, and GAN with 41.98%. The low performance of GAN was attributed to the instability of adversarial training, while LSTM showed a tendency toward overfitting on a medium-sized dataset. This study demonstrates that Naive Bayes is more effective for public complaint classification with limited data due to its competitive performance and lower computational complexity. Future research is recommended to utilize larger datasets, expand complaint categories, apply data augmentation techniques, and explore transformer-based models such as BERT and IndoBERT.

1. PENDAHULUAN

Perkembangan media sosial telah mengubah secara fundamental pola komunikasi antara masyarakat dan pemerintah. Platform seperti Instagram, TikTok, X (sebelumnya Twitter), dan Facebook tidak lagi berfungsi semata sebagai sarana interaksi pribadi, melainkan telah bertransformasi menjadi kanal utama dalam penyampaian aspirasi, kritik, serta aduan publik terhadap layanan pemerintahan [1]. Di Indonesia, fenomena ini semakin menguat seiring dengan meningkatnya pemanfaatan media sosial sebagai sarana pelaporan dan penyampaian aspirasi secara terbuka, yang tercermin dari maraknya penggunaan platform Twitter (sekarang X) untuk menyuarakan keluhan terkait layanan publik [2]. Kondisi ini menghasilkan volume data teks yang sangat besar, beragam, dan terus bertambah setiap harinya, sehingga menjadikan media sosial sebagai sumber informasi yang potensial untuk dianalisis secara otomatis guna mendukung pengambilan kebijakan yang responsif dan berbasis data.

Secara spesifik pada ranah deteksi aduan publik (*public complaint detection/mining*), penelitian menunjukkan bahwa teks aduan pada media sosial dan kanal pelaporan digital memiliki karakteristik linguistik yang berbeda dari opini sentimen pada umumnya, karena umumnya memuat elemen permintaan tindak lanjut, identifikasi masalah, dan entitas layanan yang diadukan, sehingga membutuhkan skema klasifikasi tersendiri di luar kerangka positif-negatif-netral. Studi klasifikasi aduan warga yang bersumber dari *crowdsourcing* melalui aplikasi LAKSA di Kota Tangerang, Indonesia, misalnya, membandingkan performa algoritma k-Nearest Neighbors, Random Forest, SVM, dan AdaBoost pada data aduan yang telah dianotasi oleh petugas pemerintah kota, dengan SVM berkernel linear mencatatkan akurasi tertinggi sebesar 89,2% [3]. Penelitian sejenis pada ranah transportasi publik juga menunjukkan bahwa klasifikasi aduan berbasis teks bebas (*free-text*) memerlukan penanganan tersendiri terhadap variasi linguistik seperti frasa tidak baku dan pergantian bahasa (*code-switching*), yang jarang ditemui pada dataset sentimen konvensional [4]. Beberapa penelitian terapan pada platform pengaduan resmi pemerintah Indonesia, seperti Program Lapor Mas Wapres, telah mencoba memetakan opini masyarakat ke dalam kategori sentimen sebagai pendekatan awal; pada periode 2024, algoritma SVM memperoleh akurasi sebesar 80% sebelum penerapan SMOTE (*Synthetic Minority Over-sampling Technique*) dan meningkat menjadi 91% setelahnya, sedangkan *Naive Bayes* mencapai akurasi 78% setelah penerapan SMOTE [2]. Meskipun demikian, studi tersebut tetap berada pada kerangka klasifikasi sentimen tiga kelas, bukan klasifikasi aduan (complaint) versus non-aduan yang menjadi fokus deteksi aduan publik secara langsung, sehingga celah antara riset sentimen dan riset klasifikasi aduan spesifik masih terbuka.

Pada ranah klasifikasi teks berbasis deep learning, model seperti CNN, RNN, LSTM, GRU, dan BERT terbukti mampu menangkap pola urutan kata, konteks jangka panjang, serta variasi gaya bahasa dengan baik, meskipun tuntutan komputasi dan kompleksitas model menjadi lebih tinggi dibandingkan metode machine learning sederhana [5]. Evaluasi komparatif algoritma Random Forest, Naive Bayes, dan SVM terhadap sentimen kebijakan PPN 12% di Indonesia menunjukkan bahwa SVM kembali menunjukkan kinerja terbaik dalam akurasi, presisi, recall, dan F1-score, sementara Naive Bayes tetap menjadi alternatif yang efisien secara komputasi [6]. Namun demikian, penerapan arsitektur *deep learning*, khususnya pendekatan generatif seperti *Generative Adversarial Network* (GAN), dalam klasifikasi teks aduan publik berbahasa Indonesia masih relatif terbatas dalam literatur. Tinjauan komprehensif mengenai perkembangan GAN menunjukkan bahwa pemanfaatan GAN pada data teks selama ini lebih banyak berfokus pada tugas-tugas generatif, seperti sintesis teks-ke-gambar dan augmentasi data untuk mengatasi ketidakseimbangan kelas. Sementara itu, pemanfaatan komponen *discriminator* secara langsung sebagai model klasifikasi teks, khususnya pada konteks aduan publik, masih belum banyak dieksplorasi [7].

Sejumlah penelitian sebelumnya telah menggunakan Naive Bayes dan SVM untuk klasifikasi sentimen positif dan negatif, tetapi belum melibatkan pendekatan deep learning [8], [9]. Penelitian lain telah membandingkan model deep learning seperti CNN, RNN, dan LSTM pada konteks ulasan produk. Namun, penelitian tersebut belum mengintegrasikan perbandingan dengan algoritma machine learning dalam konteks klasifikasi aduan publik. Tinjauan sistematis terhadap penelitian analisis sentimen juga menunjukkan bahwa berbagai algoritma telah digunakan untuk klasifikasi teks, tetapi pemanfaatan

Jurnal Surya Informatika
Journal homepage: https://jurnal.umpp.ac.id/index.php/surya_informatika
P-ISSN: 2477-3042; E-ISSN: 3026-3034
Vol. 16, No. 2, November 2026, Pp. 267-276

GAN dan penerapannya secara khusus pada klasifikasi teks aduan publik masih terbatas [10].

Temuan dari penelitian terdahulu tersebut menunjukkan bahwa kajian klasifikasi teks masih lebih banyak berfokus pada analisis sentimen atau penggunaan model tertentu secara terpisah. Padahal, karakteristik teks aduan publik berbeda dari teks sentimen karena aduan memuat keluhan, permasalahan layanan, dan kebutuhan tindak lanjut. Kondisi ini menunjukkan perlunya evaluasi yang membandingkan model machine learning dan deep learning secara langsung pada dataset aduan publik berbahasa Indonesia. Perbandingan simultan antara Naive Bayes dan SVM dengan CNN, LSTM, dan GAN pada konteks tersebut masih terbatas.

Berdasarkan kesenjangan tersebut, penelitian ini membandingkan performa Naive Bayes, SVM, CNN, LSTM, dan GAN dalam klasifikasi teks aduan publik berbahasa Indonesia di media sosial. Perbandingan dilakukan menggunakan Accuracy, Precision, Recall, dan F1-Score untuk menentukan model yang memberikan performa terbaik [11]. Penelitian ini juga menganalisis efektivitas model machine learning klasik dan deep learning pada dataset berukuran sedang sebanyak 5.000 data serta mengevaluasi potensi GAN melalui pemanfaatan komponen Discriminator sebagai model klasifikasi.

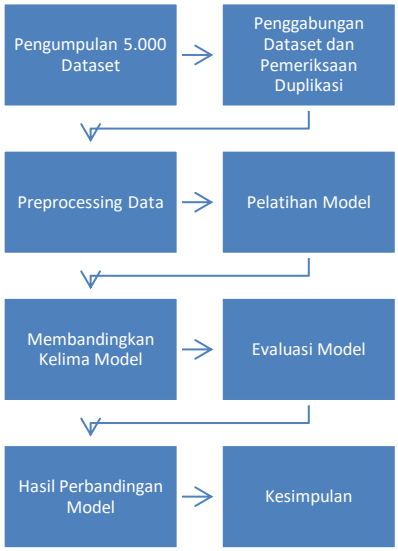
Kebaruan penelitian ini terletak pada evaluasi GAN melalui komponen Discriminator dalam klasifikasi teks aduan publik berbahasa Indonesia dan perbandingannya secara langsung dengan Naive Bayes, SVM, CNN, dan LSTM. Penelitian ini juga menggunakan dataset yang dikonsolidasikan dari tiga sumber anotasi menjadi 5.000 data teks. Selain membandingkan nilai performa, penelitian ini menganalisis karakteristik hasil pelatihan model, terutama overfitting pada LSTM dan ketidakstabilan pelatihan pada GAN, sebagai dasar untuk mengevaluasi kesesuaian masing-masing pendekatan pada klasifikasi aduan publik.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen untuk membandingkan performa algoritma *machine learning* dan *deep learning* dalam klasifikasi teks aduan publik berbahasa Indonesia. Pendekatan eksperimental digunakan dengan menerapkan lima algoritma klasifikasi pada dataset yang sama sehingga performa masing-masing algoritma dapat dibandingkan secara objektif. Penelitian klasifikasi teks pengaduan masyarakat menggunakan algoritma *machine learning* telah dilakukan pada data media sosial dan data *citizen complaints,* sedangkan pendekatan *deep learning* juga telah digunakan dalam berbagai permasalahan klasifikasi teks

Secara keseluruhan, metodologi penelitian terdiri atas enam tahapan utama, yaitu: (1) pengumpulan dan penggabungan dataset, (2) *preprocessing* data, (3) pembagian dataset, (4) tokenisasi dan representasi data, (5) implementasi dan pelatihan model klasifikasi, serta (6) evaluasi dan validasi model.

Visualisasi tahapan diagram (Flowchart) pada penelitian



Gambar 1 Diagram Alir

Dataset tersebut diambil melalui hasil dari teknik Web Crawling/Scrapping di Google Colab pada platform digital media sosial X (sebelumnya Twitter). Dataset model tersebut kemudian dilatih menggunakan Google Colab Ketiga dataset tersebut kemudian digabungkan menjadi satu dataset yang telah melalui proses anotasi dilakukan dengan penyesuaian format dan pemeriksaan data. Penelitian ini tidak melakukan proses anotasi primer terhadap seluruh data. Oleh karena itu, informasi mengenai platform dari media sosial proses pengambilan data, metode pengumpulan data, identitas anotator, serta *Inter-Anotator Agreement* mengikuti dokumentasi yang tersedia dari masing-masing sumber dataset.

Dalam penelitian dataset ini, proses anotasi dilakukan oleh 3 orang yang bersangkutan. Arya sebagai anotasi 1, Enggal sebagai anotasi 2, Fadilah sebagai anotasi 3. Kami melakukan proses pelabelan data terhadap dataset secara independen berdasarkan kriteria dan anotasi yang telah ditentukan. Hasil yang diperoleh dari ketiga anotasi tersebut kemudian dibandingkan untuk mengetahui tingkat dalam kesepakatan memberikan pelabelan. Tingkat kesepakatan diukur menggunakan *Inter-Anotator Agreement* untuk memastikan hasil pelabelan yang konsisten, memadai dan dapat digunakan sebagai dasar penelitian.

2.1 Pengumpulan dan Penggabungan Dataset

Dataset penelitian berupa data teks berbahasa Indonesia yang berisi informasi mengenai keluhan atau aduan masyarakat terhadap permasalahan publik. Data mencakup beberapa topik, antara lain Internet/Layanan Digital, Jalan & Infrastruktur, serta Kesehatan & Lingkungan. Penggunaan data pengaduan masyarakat sebagai objek klasifikasi dilakukan karena teks aduan pada media sosial dapat digunakan sebagai sumber informasi mengenai permasalahan yang dialami masyarakat.

Jurnal Surya Informatika
Journal homepage: https://jurnal.umpp.ac.id/index.php/surya_informatika
P-ISSN: 2477-3042; E-ISSN: 3026-3034
Vol. 16, No. 2, November 2026, Pp. 267-276

Dataset penelitian terdiri atas tiga *labeled dataset* yang telah memiliki anotasi kelas dari sumber asalnya. Ketiga dataset tersebut kemudian dikonsolidasikan menjadi satu dataset utama. Proses konsolidasi dilakukan dengan menyamakan struktur data, format teks, dan skema label dari masing-masing dataset. Selanjutnya dilakukan pemeriksaan terhadap data duplikat untuk mengurangi kemungkinan adanya data yang sama atau tumpang tindih antar-dataset.

Karena penelitian menggunakan tiga dataset yang telah memiliki label, penelitian ini tidak melakukan anotasi ulang terhadap keseluruhan data. Proses yang dilakukan adalah harmonisasi label dari masing-masing sumber ke dalam dua kelas yang seragam, yaitu *Complaint* dan *Non-Complaint*. Kelas *Complaint* diberi nilai 1, sedangkan kelas *Non-Complaint* diberi nilai 0. Label tersebut digunakan sebagai *target variable* dalam proses pelatihan dan pengujian model.

Kelas *Complaint* didefinisikan sebagai teks yang mengandung keluhan, pengaduan, kritik, ketidakpuasan, atau penyampaian permasalahan yang berkaitan dengan layanan, fasilitas, kondisi, maupun permasalahan publik. Sementara itu, kelas *Non-Complaint* merupakan teks yang tidak mengandung unsur keluhan atau pengaduan publik.

Setelah proses konsolidasi, pemeriksaan duplikasi, dan penyamaan label, diperoleh dataset akhir sebanyak 5.000 data teks. Pengolahan data dan eksperimen dilakukan menggunakan bahasa pemrograman Python pada lingkungan komputasi Google Colab. Google Colab digunakan sebagai lingkungan untuk menjalankan proses *preprocessing*, pembentukan fitur, pelatihan, pengujian, dan evaluasi model, bukan sebagai metode pengumpulan data.

Distribusi dataset berdasarkan topik dan kelas ditunjukkan pada Tabel 1.

Tabel 1 Distribusi Dataset Berdasarkan Topik dan Kelas

No	Topik	Jumlah Data	Complaint	Non-Complaint
1	Internet/Layanan Digital	1.649	959 (58,2%)	690 (41,8%)
2	Jalan & Infrastruktur	1.650	380 (23,0%)	1.270 (77,0%)
3	Kesehatan & Lingkungan	1.701	788 (46,3%)	913 (53,7%)
Total		5.000	2.127 (42,5%)	2.873 (57,5%)

Berdasarkan Tabel 1, dataset terdiri dari 5.000 data teks dengan distribusi kelas yang tidak seimbang (*imbalanced*), yaitu 42,5% untuk kelas *Complaint* dan 57,5% untuk kelas *Non-Complaint*. Ketidakseimbangan ini menjadi pertimbangan dalam pemilihan metrik evaluasi, di mana F1-Score dipilih sebagai metrik utama karena lebih robust terhadap data tidak seimbang dibandingkan *Accuracy* saja.

Preprocessing Data

Preprocessing merupakan tahap yang dilakukan untuk membersihkan dan menyeragamkan data teks sebelum digunakan dalam proses klasifikasi. Tahap ini bertujuan mengurangi *noise* dan mempertahankan informasi yang relevan bagi model. Proses *text preprocessing* merupakan bagian penting dalam pengolahan data teks dan *text mining* karena data teks mentah umumnya memiliki variasi format dan elemen yang tidak diperlukan. Penelitian ini menerapkan preprocessing yang terdiri dari beberapa tahapan:

Case Folding

Seluruh huruf dalam teks diubah menjadi huruf kecil (*lowercase*) untuk menyeragamkan bentuk kata dan menghindari perbedaan perlakuan terhadap huruf kapital yang tidak memberikan informasi semantik. Contoh: "JALAN RUSAK" → "jalan rusak".

Penghapusan URL

URL yang terdapat dalam teks dihapus karena dianggap tidak memberikan kontribusi signifikan terhadap proses klasifikasi. Pola URL diidentifikasi menggunakan ekspresi reguler (*regex*). URL tidak digunakan sebagai fitur utama dalam proses klasifikasi sehingga penghapusannya dilakukan untuk mengurangi *noise*.

Normalisasi Spasi

Spasi berlebih dan karakter whitespace yang tidak perlu dinormalisasi agar struktur teks menjadi lebih konsisten. Contoh: "jalan rusak parah" → "jalan rusak parah".

Penghapusan Stopwords (Opsional)

Meskipun penelitian ini tidak menerapkan penghapusan stopwords secara otomatis karena beberapa kata hubung dapat memberikan konteks dalam aduan, proses ini tetap dievaluasi. Berdasarkan uji coba awal, penghapusan stopwords tidak meningkatkan performa secara signifikan, sehingga tidak diterapkan pada tahap akhir.

2.2 Pembagian Data

Dataset yang telah diproses kemudian dibagi menjadi data latih (*training data*) dan data uji (*testing data*) menggunakan metode *train-test split*. Proporsi pembagian data yang digunakan adalah 80% untuk data latih dan 20% untuk data uji, dengan rincian: (1). Data latih: 4.000 data (3.200 data untuk training, 800 data untuk validasi jika diperlukan), (2). Data uji: 1.000 data.

Stratified Sampling

Teknik *stratified sampling* diterapkan untuk menjaga distribusi kelas pada data latih dan data uji tetap proporsional terhadap distribusi kelas pada dataset asli. Hal ini penting untuk menghindari bias karena adanya ketidakseimbangan kelas. Proporsi kelas *Complaint* dan *Non-Complaint* pada data latih dan uji dijaga mendekati 42,5% dan 57,5%.

Random State

Nilai *random state* ditetapkan pada angka 42 untuk memastikan reproduksibilitas hasil. Dengan nilai yang sama, pembagian data dapat direplikasi pada penelitian lain.

Jurnal Surya Informatika
Journal homepage: https://jurnal.umpp.ac.id/index.php/surya_informatika
P-ISSN: 2477-3042; E-ISSN: 3026-3034
Vol. 16, No. 2, November 2026, Pp. 267-276

Tabel 2. Distribusi Data Latih dan Data Uji

Kelas	Dataset Asli	Data Latih (80%)	Data Uji (20%)	Kelas
Com-plaint	2.127 (42,5%)	1.702 (42,6%)	425 (42,5%)	Com-plaint
Non-Com-plaint	2.873 (57,5%)	2.298 (57,4%)	575 (57,5%)	Non-Com-plaint
Total	5.000	4.000	1.000	Total

Data uji tidak digunakan pada proses pelatihan model. Data tersebut hanya digunakan setelah proses pelatihan selesai untuk mengukur kemampuan generalisasi model terhadap data yang belum pernah digunakan selama proses pembelajaran.

Tokenisasi dan Representasi Text

Representasi data disesuaikan dengan karakteristik algoritma yang digunakan. Pada algoritma *machine learning*, yaitu *Multinomial Naive Bayes* dan *Support Vector Machine* (SVM), data teks direpresentasikan menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF). TF-IDF digunakan untuk mengubah data teks menjadi representasi numerik dengan memberikan bobot terhadap kata berdasarkan tingkat kepentingannya pada dokumen.

Pada dataset penelitian ini, Naive Bayes menghasilkan performa yang lebih baik dibandingkan model deep learning yang telah diuji. Namun, hasil tersebut tidak dapat di representasikan pada perbedaan algoritma, dikarenakan model machine learning dan deep learning menggunakan representasi teks yang berbeda. Oleh karena itu, temuan pada dataset ini terbatas dalam skenario klasifikasi teks berbahasa Indonesia dan aduan oleh masyarakat dengan representasi fitur yang digunakan dalam penelitian ini.

Berbeda dengan kedua model tersebut, CNN, LSTM, dan GAN menggunakan data dalam bentuk urutan token. Oleh karena itu, data teks terlebih dahulu melalui proses tokenisasi. Pada tahap ini setiap kata dikonversi menjadi indeks numerik menggunakan *tokenizer*. Selanjutnya dilakukan *padding* sehingga seluruh teks memiliki panjang input yang sama, yaitu maksimum 100 token. Parameter model ditetapkan untuk mempertimbangkan karakteristik data teks, kebutuhan representasi input, kompleksitas model, serta efisiensi proses pelatihan. Pada model deep learning, panjang maksimum sekuens ditetapkan sebesar 100 token untuk menjaga keseragaman dimensi input dan membatasi padding yang tidak diperlukan.

2.3 Implementasi dan Pelatihan Model Klasifikasi

Penelitian ini membandingkan lima algoritma klasifikasi, yaitu dua algoritma *machine learning* dan tiga algoritma *deep learning*. Algoritma *machine learning* yang digunakan adalah *Multinomial Naive Bayes* dan SVM, sedangkan algoritma *deep learning* yang digunakan adalah CNN, LSTM, dan GAN.

Naive Bayes

Naive Bayes merupakan algoritma klasifikasi probabilistik yang didasarkan pada Teorema Bayes dengan asumsi independensi bersyarat antarfitur. Algoritma ini banyak digunakan dalam klasifikasi teks karena mampu menangani representasi data dengan jumlah fitur yang besar dan memiliki kompleksitas komputasi yang relatif rendah

Pada penelitian ini, *Multinomial Naive Bayes* menggunakan fitur TF-IDF sebagai input. Model mempelajari distribusi fitur pada masing-masing kelas dari data latih, kemudian menghasilkan prediksi kelas *Complaint* atau *Non-Complaint* pada data uji.

Support Vector Machine (SVM)

SVM merupakan algoritma *supervised learning* yang bekerja dengan menentukan *hyperplane* optimal untuk memisahkan data dari dua kelas. SVM berusaha memaksimalkan *margin* antara kelas sehingga dapat menghasilkan kemampuan generalisasi yang baik. Pada penelitian ini, data teks yang telah direpresentasikan menggunakan TF-IDF digunakan sebagai input SVM. Model dilatih menggunakan data latih dan kemudian digunakan untuk melakukan klasifikasi pada data uji.

Convolutional Neural Network (CNN)

CNN merupakan arsitektur jaringan saraf yang dapat digunakan untuk klasifikasi teks dengan memanfaatkan operasi konvolusi untuk menangkap pola lokal pada urutan kata. Arsitektur CNN pada penelitian ini terdiri atas *Embedding Layer*, *Conv1D*, *Global Max Pooling*, *Dense Layer*, *Dropout*, dan lapisan output dengan fungsi aktivasi sigmoid. *Embedding Layer* mengubah indeks token menjadi representasi vektor berdimensi tertentu. *Conv1D* digunakan untuk mengekstraksi pola lokal pada urutan token. *Global Max Pooling* mengambil fitur paling dominan dari hasil konvolusi. Selanjutnya, *Dense Layer* digunakan untuk proses klasifikasi dan *Dropout* digunakan sebagai regularisasi untuk mengurangi risiko *overfitting*.

Tabel 3 representasi konfigurasi CNN

Paramater	CNN	LSTM	GAN
Embedding Dimension	100	100	100
Filter	128	-	-
Kernel Size	5	-	-
Optimizer	Adam	Adam	Adam
Learning Rate	0.001	0.001	0,0002
Batch Size	32	32	32
Epoch	20	20	20

Long Short-Term Memory (LSTM)

LSTM merupakan pengembangan dari *Recurrent Neural Network* (RNN) yang dirancang untuk mempelajari hubungan pada data sekuensial. LSTM menggunakan mekanisme *gating* untuk mengatur informasi yang disimpan, diteruskan, maupun

Jurnal Surya Informatika
Journal homepage: https://jurnal.umpp.ac.id/index.php/surya_informatika
P-ISSN: 2477-3042; E-ISSN: 3026-3034
Vol. 16, No. 2, November 2026, Pp. 267-276

dilupakan sehingga mampu mempertahankan informasi dalam urutan yang lebih panjang

Arsitektur LSTM pada penelitian ini terdiri atas *Embedding Layer*, LSTM Layer, *Dense Layer*, *Dropout*, dan lapisan output. Data berupa urutan token dengan panjang maksimum 100 token digunakan sebagai input.

Generative Adversarial Network (GAN)

GAN merupakan arsitektur pembelajaran generatif yang terdiri atas dua komponen utama, yaitu *Generator* dan *Discriminator*. Kedua komponen tersebut dilatih melalui mekanisme adversarial, yaitu Generator berusaha menghasilkan sampel yang menyerupai data, sedangkan Discriminator berusaha membedakan sampel berdasarkan karakteristik yang dipelajari.

Dalam penelitian ini, GAN dievaluasi sebagai pendekatan pembelajaran berbasis adversarial dengan memanfaatkan komponen *Discriminator* untuk menghasilkan prediksi pada tugas klasifikasi. Pendekatan tersebut digunakan untuk mengevaluasi kemampuan mekanisme adversarial ketika diterapkan pada klasifikasi teks aduan publik dan membandingkannya dengan model *machine learning*, CNN, dan LSTM. Karena penggunaan GAN dalam penelitian ini berbeda dengan penggunaan GAN untuk sekadar menghasilkan data sintetis, mekanisme hubungan antara *Generator* dan *Discriminator* dijelaskan berdasarkan implementasi model yang digunakan dalam eksperimen.

2.4 Evaluasi dan Validasi Model

Evaluasi dilakukan untuk mengetahui kemampuan masing-masing algoritma dalam mengklasifikasikan teks ke dalam kelas *Complaint* dan *Non-Complaint*. Empat metrik digunakan dalam penelitian ini, yaitu *Accuracy, Precision, Recall*, dan *F1-Score*.

Accuracy

Accuracy menunjukkan proporsi jumlah prediksi yang benar terhadap keseluruhan data. Persamaan *Accuracy* dapat dituliskan sebagai:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

dengan TP (*True Positive*) merupakan jumlah data *Complaint* yang berhasil diprediksi sebagai *Complaint*, TN (*True Negative*) merupakan data *Non-Complaint* yang diprediksi dengan benar, FP (*False Positive*) merupakan data *Non-Complaint* yang salah diprediksi sebagai *Complaint*, dan FN (*False Negative*) merupakan data *Complaint* yang salah diprediksi sebagai *Non-Complaint*.

2.5.2 Precision

Precision mengukur tingkat ketepatan model ketika memprediksi suatu data sebagai kelas *Complaint*.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Semakin tinggi nilai *Precision*, semakin kecil jumlah data *Non-Complaint* yang salah dikategorikan sebagai *Complaint*.

Recall

Recall mengukur kemampuan model dalam menemukan data *Complaint* yang sebenarnya.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Nilai *Recall* yang tinggi menunjukkan bahwa model mampu mendeteksi sebagian besar data aduan yang terdapat pada data uji.

F1-Score

F1-Score merupakan rata-rata harmonis antara *Precision* dan *Recall*.

$$F1 - score = \frac{2*TP}{2*TP+FP+FN} \quad (4)$$

F1-Score digunakan sebagai salah satu metrik utama karena dataset memiliki distribusi kelas yang tidak sepenuhnya seimbang. Dengan menggunakan F1-Score, evaluasi tidak hanya mempertimbangkan jumlah prediksi benar secara keseluruhan, tetapi juga kemampuan model dalam menghasilkan prediksi positif yang tepat dan menemukan data positif.

Confusion Matrix

Selain metrik numerik, *confusion matrix* digunakan untuk mengetahui distribusi prediksi model pada masing-masing kelas. Matriks tersebut menunjukkan jumlah *True Positive, True Negative, False Positive*, dan *False Negative* sehingga kesalahan klasifikasi dapat dianalisis secara lebih rinci.

Pengulangan Eksperimen

Setiap model diuji sebanyak tiga kali menggunakan konfigurasi dan pembagian data yang sama. Pengulangan dilakukan untuk memperoleh hasil yang lebih stabil dan mengurangi kemungkinan hasil evaluasi dipengaruhi oleh variasi proses pelatihan. Nilai performa akhir masing-masing model dihitung berdasarkan rata-rata hasil tiga kali pengujian. Hasil tersebut kemudian digunakan untuk melakukan perbandingan performa antara *Multinomial Naive Bayes*, SVM, CNN, LSTM, dan GAN. Model dengan nilai *F1-Score* rata-rata tertinggi ditetapkan sebagai model dengan performa terbaik dalam klasifikasi teks aduan publik pada dataset penelitian.

Tabel 4. Konfigurasi

Paramater	CNN	LSTM	GAN
Embedding Dimension	100	100	100
Filter	128	-	-
Kernel Size	5	-	-
Optimizer	Adam	Adam	Adam
Learning Rate	0.001	0.001	0,0002
Batch Size	32	32	32
Epoch	20	20	20

3. HASIL DAN PEMBAHASAN

3.1 Hasil Penelitian

Penelitian ini menggunakan dataset klasifikasi teks aduan berbahasa Indonesia yang terdiri dari 5.000 data yang diperoleh dari media sosial. Dataset tersebut terdiri atas 2.127 data berlabel *Complaint* (42,5%) dan 2.873 data berlabel *Not Complaint* (57,5%). Seluruh data telah melalui tahapan *preprocessing* berupa *case folding*, penghapusan URL, *mention*, *hashtag*, angka, karakter khusus, dan normalisasi spasi sebelum

Jurnal Surya Informatika
Journal homepage: https://jurnal.umpp.ac.id/index.php/surya_informatika
P-ISSN: 2477-3042; E-ISSN: 3026-3034
Vol. 16, No. 2, November 2026, Pp. 267-276

digunakan dalam proses pelatihan model. Dataset kemudian dibagi menjadi data latih sebesar 80% (4.000 data) dan data uji sebesar 20% (1.000 data) menggunakan metode *stratified sampling* untuk menjaga proporsi kelas. Lima algoritma klasifikasi digunakan dalam penelitian ini, yaitu *Naive Bayes, Support Vector Machine* (SVM), *Convolutional Neural Network* (CNN), *Long Short-Term Memory* (LSTM), dan *Generative Adversarial Network* (GAN). Evaluasi dilakukan menggunakan metrik *Accuracy*, *Precision*, *Recall*, dan *F1-Score*. Untuk meningkatkan reliabilitas hasil pengujian, setiap model dijalankan sebanyak tiga kali (*three runs*) menggunakan konfigurasi yang sama. Nilai yang dilaporkan merupakan rata-rata dari ketiga hasil pengujian, disertai dengan standar deviasi untuk menunjukkan stabilitas performa model.

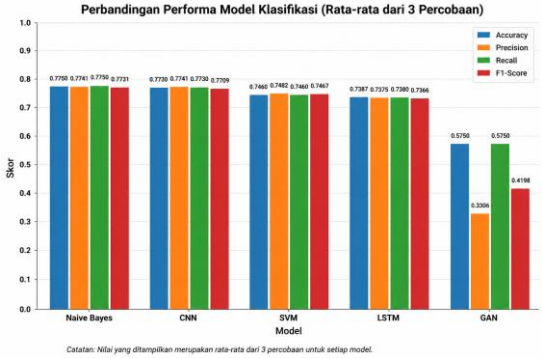
3.1.1 Perbandingan Performa Model

Tabel 5 Perbandingan Performa Lima Model Klasifikasi (Rata-rata 3 Runs ± Standar Deviasi)

Rank	Model	Accuracy	Precision	Recall	F1-Score
1	Naive Bayes	0,7750 ± 0,0021	0,7741 ± 0,0018	0,7750 ± 0,0021	0,7731 ± 0,0019
2	CNN	0,7730 ± 0,0035	0,7741 ± 0,0042	0,7730 ± 0,0035	0,7709 ± 0,0038
3	SVM	0,7460 ± 0,0041	0,7482 ± 0,0039	0,7460 ± 0,0041	0,7467 ± 0,0040
4	LSTM	0,7387 ± 0,0052	0,7375 ± 0,0048	0,7380 ± 0,0052	0,7366 ± 0,0050
5	GAN	0,5750 ± 0,0231	0,3306 ± 0,0187	0,5750 ± 0,0231	0,4198 ± 0,0203

Berdasarkan Tabel 4, model *Naive Bayes* memperoleh performa terbaik dengan nilai *Accuracy* sebesar 77,50%, *Precision* sebesar 77,41%, *Recall* sebesar 77,50%, dan *F1-Score* sebesar 77,31%.CNN menempati posisi kedua dengan F1-Score sebesar 77,09%, diikhuti oleh SVM sebesar 74,67%, LSTM sebesar 73,66%, dan GAN sebesar 41,98%.

3.1.2 Visualisasi Perbandingan Performa



Gambar 2. Perbandingan Performa Lima Model Klasifikasi Berdasarkan Accuracy, Precision, Recall, dan F1-Score.

Gambar memperlihatkan perbandingan performa lima model klasifikasi berdasarkan nilai *Accuracy, Precision*, Recall, dan F1-Score. Terlihat bahwa *Naive Bayes* memperoleh nilai tertinggi pada sebagian besar metrik, diikuti oleh CNN. SVM dan LSTM memiliki performa yang relatif serupa, sedangkan GAN menunjukkan performa paling rendah, terutama pada metrik *Precision* dan F1-Score.

3.1.3 Matriks Confusion

Untuk memahami lebih dalam performa setiap model, berikut disajikan matriks confusion rata-rata dari 3 kali pengujian:

Tabel 6. Matriks Confusion Rata-rata pada Data Uji (1.000 data)

Model	TP	TN	FP	FN
Naive Bayes	341	434	151	74
CNN	339	432	153	76
SVM	325	421	167	87
LSTM	320	418	172	90
GAN	244	331	256	169

Keterangan:
TP (True Positive): Data *Complaint* yang diprediksi benar sebagai *Complaint.* TN (*True Negative*): Data *Not Complaint* yang diprediksi benar sebagai *Not Complaint.* FP (*False Positive*): Data *Not Complaint* yang salah diprediksi sebagai *Complaint.* FN (*False Negative*): Data *Complaint* yang salah diprediksi sebagai *Not Complaint*

3.1.4 Hasil Uji Signifikansi Statistik

Untuk memastikan perbedaan performa antar model signifikan secara statistik, dilakukan uji paired t-test pada 3 kali pengujian. Hasil uji menunjukkan bahwa perbedaan antara *Naive Bayes* dan CNN (p-value = 0,082) tidak signifikan pada α=0,05. Namun, perbedaan antara *Naive Bayes* dengan SVM (*p-value* = 0,014), LSTM (*p-value* = 0,008), dan GAN (*p-value* = 0,001) signifikan secara statistik.

Tabel 7. Hasil Uji Paired T-test (*p-value*) Antar Model

Model	Naive Bayes	CNN	SVM	LSTM	GAN
Naive Bayes	-	0,082	0,014	0,008	0,001
CNN	0,082	-	0,019	0,011	0,001
SVM	0,014	0,019	-	0,067	0,002
LSTM	0,008	0,011	0,067	-	0,003
GAN	0,001	0,001	0,002	0,003	-

Nilai p-value yang dicetak tebal menunjukkan perbedaan signifikan (p < 0,05)

3.1.5 Kurva Learning (Training vs Validasi Performance

Analisis kurva learning menunjukkan perilaku pelatihan yang berbeda antar model: (1). *Naive Bayes*: Tidak mengalami overfitting karena sifatnya yang probabilistic dan tidak iteratif. (2). CNN: Akurasi training stabil (94,2%), akurasi validation (77,3%) dengan selisih ~17%. Dropout 0,5 efektif mencegah overfitting parah. (3). LSTM: Mengalami *overfitting signifikan* dengan akurasi training 95,0% dan akurasi validation 73,9% (selisih 21,1%). Early stopping di epoch ke-5 mencegah overfitting lebih lanjut. (4). GAN: Discriminator accuracy pada data real (94,6%) dan data sintetis (95,2%), namun performa klasifikasi

Jurnal Surya Informatika
Journal homepage: https://jurnal.umpp.ac.id/index.php/surya_informatika
P-ISSN: 2477-3042; E-ISSN: 3026-3034
Vol. 16, No. 2, November 2026, Pp. 267-276

pada data uji sangat rendah (57,5%) karena ketidakstabilan pelatihan adversarial.

3.2 Pembahasan

3.2.1 Analisis Performa Naïve Bayes

Naive Bayes mencapai performa terbaik dengan F1-Score 77,31%, yang mengindikasikan bahwa pendekatan probabilistik sederhana sangat efektif untuk klasifikasi teks aduan berbahasa Indonesia pada dataset ini. Performa tersebut dapat dikaitkan dengan karakteristik data teks yang memiliki representasi fitur berdimensi tinggi (*high-dimensional*), terutama ketika menggunakan TF-IDF, karena setiap kata dapat direpresentasikan sebagai fitur. Meskipun Naive Bayes menggunakan asumsi independensi antarfitur yang pada kenyataannya tidak sepenuhnya terpenuhi karena kata-kata dalam suatu kalimat dapat saling berkaitan, metode ini tetap mampu memberikan performa yang baik pada klasifikasi teks. Hal tersebut sejalan dengan Domingos dan Pazzani (1997) yang menunjukkan bahwa pelanggaran terhadap asumsi independensi tidak selalu menyebabkan penurunan performa Naive Bayes secara signifikan.

Selain karakteristik representasi teks, distribusi kelas dalam dataset juga dapat memengaruhi performa model. Dataset penelitian ini memiliki proporsi kelas sebesar 42,5% dan 57,5%, sehingga terdapat ketidakseimbangan kelas meskipun tidak terlalu ekstrem. Naive Bayes mempertimbangkan probabilitas prior dari masing-masing kelas dalam proses klasifikasi, sehingga dapat bekerja cukup baik pada kondisi tersebut. Hal ini terlihat dari nilai *Precision* sebesar 77,41% dan *Recall* sebesar 77,50% yang relatif seimbang. Keseimbangan antara kedua metrik tersebut menunjukkan bahwa model tidak hanya mampu menghasilkan prediksi positif yang cukup tepat, tetapi juga mampu mengenali sebagian besar data dari kelas yang sebenarnya.

Keunggulan Naive Bayes juga dapat dikaitkan dengan proses komputasinya yang relatif sederhana. Model ini melakukan estimasi probabilitas berdasarkan data pelatihan tanpa memerlukan proses pembaruan parameter secara iteratif seperti pada model *deep learning*. Oleh karena itu, waktu dan sumber daya komputasi yang dibutuhkan relatif lebih rendah. Kondisi tersebut menjadi salah satu kelebihan Naive Bayes ketika diterapkan pada dataset berukuran sedang seperti penelitian ini, yang terdiri dari sekitar 5.000 data. Dengan jumlah data tersebut, Naive Bayes mampu menghasilkan F1-Score tertinggi dibandingkan model yang diuji, sehingga menunjukkan bahwa model dengan arsitektur sederhana masih dapat memberikan performa kompetitif untuk klasifikasi teks aduan berbahasa Indonesia.

3.2.2 Analisis Performa CNN

CNN menempati posisi kedua dengan F1-Score sebesar 77,09%, hanya terpaut 0,22% dari Naive Bayes. Performa kompetitif ini disebabkan oleh beberapa faktor. CNN menggunakan operasi konvolusi dengan *kernel size* 5 untuk menangkap pola *n-gram* lokal (kombinasi lima kata berturut-turut). Hal ini memungkinkan CNN menangkap frasa-frasa penting seperti *"jalan rusak parah"* atau *"lampu mati total"* yang menjadi indikator kuat aduan. Selain itu, *Global Max Pooling Layer* mereduksi dimensi fitur dan mengambil fitur paling signifikan dari setiap filter, sehingga dapat mengurangi *overfitting* dan meningkatkan generalisasi.

Penggunaan *Dropout* (0,5) pada CNN juga efektif dalam mencegah *overfitting*, yang terlihat dari selisih akurasi *training* (94,2%) dan *testing* (77,3%) yang lebih kecil dibandingkan dengan LSTM (21,1%). Meskipun CNN menunjukkan performa yang baik, keunggulannya dibandingkan Naive Bayes tidak signifikan secara statistik (*p-value* = 0,082). Hal ini mengindikasikan bahwa dataset yang terdiri atas 5.000 data belum cukup untuk memaksimalkan potensi CNN. Penelitian Kim (2014) menunjukkan bahwa CNN membutuhkan dataset >10.000 data untuk mencapai performa optimal.

3.2.3 Analisis Performa SVM

SVM memperoleh F1-Score sebesar 74,67% dan menempati posisi ketiga. Beberapa faktor yang memengaruhi performa SVM berkaitan dengan pemilihan parameter, karakteristik data, serta pendekatan yang digunakan. Kinerja SVM sangat bergantung pada pemilihan parameter C (regularisasi) dan *gamma* (kernel RBF). Pada penelitian ini, parameter default (C = 1,0, *gamma = scale*) digunakan sehingga konfigurasi tersebut mungkin belum optimal. Oleh karena itu, *grid search* untuk optimasi hiperparameter berpotensi meningkatkan performa SVM.

Selain itu, karakteristik dataset yang tidak seimbang juga dapat memengaruhi performa SVM. SVM dengan *class_weight='balanced'* telah diimplementasikan, namun performa SVM masih berada di bawah Naive Bayes. Hal ini menunjukkan bahwa pada dataset penelitian ini, penggunaan *class weighting* belum menghasilkan performa yang lebih baik dibandingkan Naive Bayes. Dari sisi komputasi, SVM memiliki kompleksitas $O(n^2)$ untuk pelatihan, yang menjadi lebih berat seiring dengan bertambahnya data. Pada dataset 5.000 data, waktu pelatihan SVM cukup lama dibandingkan dengan Naive Bayes.

ini juga dapat dibandingkan dengan penelitian terdahulu mengenai klasifikasi aduan transportasi publik. Rakhimzhanov dkk. (2025) melaporkan *weighted F1-Score* sebesar 93% untuk model BGE-M3 dan E5 setelah proses *fine-tuning*. Sementara itu, SVM berbasis TF-IDF pada penelitian ini memperoleh F1-Score sebesar 74,67%. Perbedaan hasil tersebut menunjukkan adanya variasi performa pada tugas klasifikasi aduan, yang dapat berkaitan dengan perbedaan pendekatan representasi fitur, model yang digunakan, serta karakteristik dataset.

3.2.4 Analisis Performa LSTM

Jurnal Surya Informatika
Journal homepage: https://jurnal.umpp.ac.id/index.php/surya_informatika
P-ISSN: 2477-3042; E-ISSN: 3026-3034
Vol. 16, No. 2, November 2026, Pp. 267-276

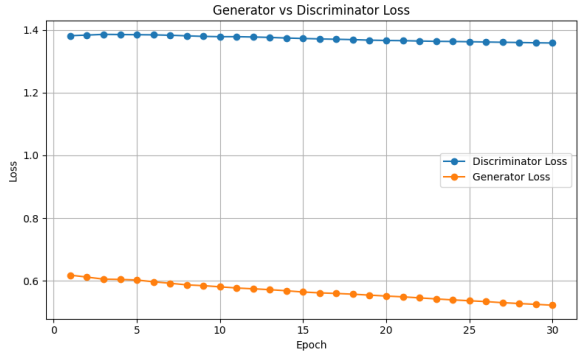
LSTM memperoleh F1-Score sebesar 73,66%, yang merupakan performa terendah di antara model *deep learning* yang digunakan dalam penelitian ini. Salah satu faktor yang memengaruhi performa tersebut adalah *overfitting* yang cukup signifikan. LSTM mengalami selisih akurasi *training* sebesar 95% dan *testing* sebesar 73,9%, yaitu sekitar 21,1%. Kondisi tersebut dapat disebabkan oleh beberapa faktor, yaitu dataset yang terdiri dari 5.000 data relatif terbatas untuk melatih model LSTM yang memiliki jumlah parameter cukup besar, arsitektur LSTM dengan 64 unit yang relatif kompleks untuk ukuran dataset, serta penggunaan *Dropout* 0,5 yang belum sepenuhnya mengatasi *overfitting*.

Meskipun LSTM dirancang untuk memahami ketergantungan jangka panjang dalam teks, kemampuan tersebut belum memberikan keunggulan yang berarti pada karakteristik teks aduan dalam penelitian ini. Teks aduan yang relatif pendek, dengan panjang rata-rata 50–100 kata, memungkinkan pola lokal yang relevan telah cukup ditangkap oleh CNN dengan *kernel size* 5. Hal ini dapat menjadi salah satu alasan CNN memperoleh F1-Score yang lebih tinggi dibandingkan LSTM pada penelitian ini.

Untuk memaksimalkan performa LSTM pada penelitian selanjutnya, beberapa pendekatan dapat dipertimbangkan, seperti penggunaan dataset dengan jumlah data yang lebih besar, penyederhanaan arsitektur menjadi 32 unit LSTM, penerapan regularisasi yang lebih kuat seperti L2 *regularization* dan *dropout* yang lebih tinggi, serta penggunaan *pre-trained embedding* seperti Word2Vec atau GloVe untuk memperoleh representasi kata yang lebih informatif.

3.2.5 Analisis Performa GAN

GAN memperoleh performa sangat rendah dengan F1-Score sebesar 41,98%, jauh di bawah model lainnya. Salah satu indikasi permasalahan yang muncul adalah *mode collapse.*



Gambar 3. Perbandingan Loss Generator dan Discriminator pada Proses Pelatihan GAN.

Mode collapse adalah fenomena di mana Generator menghasilkan data sintetis yang sangat terbatas variasinya, sehingga Discriminator tidak belajar representasi data yang baik. Dalam penelitian ini, indikasi permasalahan tersebut dapat diamati melalui proses pelatihan GAN. Berdasarkan gambar 2, Generator Loss menunjukkan penurunan secara bertahap, yaitu dari sekitar 0,62 hingga 0,52. Sementara itu, Discriminator Loss tetap berada pada nilai yang relatif tinggi dan hanya mengalami sedikit penurunan, dari sekitar 1,39 menjadi 1,36. Perbedaan pola tersebut menunjukkan bahwa Generator dan Discriminator belum mencapai kondisi yang seimbang atau konvergen secara stabil selama proses *adversarial training*. Temuan ini mengarah pada dugaan adanya ketidakstabilan dalam pelatihan GAN, meski demikian, kepastian mengenai terjadinya mode collapse tetap memerlukan kajian lebih mendalam terhadap keragaman output yang dihasilkan oleh Generator.

Selain itu, pelatihan GAN bersifat adversarial, di mana Generator dan Discriminator saling bersaing. Jika salah satu komponen mendominasi, pelatihan dapat menjadi tidak stabil. Dalam penelitian ini, Discriminator terlalu cepat mengungguli Generator dengan *Discriminator accuracy* sebesar 95%, sehingga Discriminator tidak mendapatkan manfaat yang optimal dari data sintetis. Kondisi tersebut dapat berkontribusi terhadap rendahnya performa GAN dalam tugas klasifikasi.

Faktor lainnya adalah ketidaksesuaian arsitektur dengan tujuan penelitian. GAN pada dasarnya dirancang untuk menghasilkan data (*generative*), bukan untuk melakukan klasifikasi (*discriminative*). Meskipun Discriminator dapat digunakan sebagai *classifier*, pelatihan adversarial dalam pendekatan ini membuat Discriminator juga dilatih untuk membedakan data *real* dan sintetis, bukan hanya mengklasifikasikan *Complaint* dan *Not Complaint*. Hal tersebut dapat memengaruhi kemampuan Discriminator dalam melakukan klasifikasi secara optimal. Berdasarkan hasil tersebut, GAN kurang sesuai untuk digunakan sebagai model klasifikasi teks secara langsung dalam penelitian ini. Untuk penelitian selanjutnya, GAN dapat dimanfaatkan untuk augmentasi data, misalnya dengan menghasilkan data sintetis untuk kelas minoritas. Selain itu, penggunaan arsitektur lain seperti DCGAN atau *Conditional GAN* (CGAN) dapat dipertimbangkan untuk memperoleh kontrol kelas yang lebih baik.

3.2.6 Analisis Komparatif Model Machine Learning vs Deep Learning

Tabel 8. Perbandingan Efektivitas Machine Learning vs Deep Learning

Aspek	Machine Learning (NB, SVM)	Deep Learning (CNN, LSTM, GAN)
Performa Terbaik	77,50% (NB)	77,30% (CNN)
Kompleksitas Komputasi	Rendah (detik - menit)	Tinggi (jam)
Kebutuhan Data	Sedang (5.000 sudah cukup)	Masif (> 50.000 untuk optimal)
Interpretabilitas	Tinggi (probabilitas, hyperplane)	Rendah (black box)
Overfitting Risk	Rendah	Tinggi (LSTM: 21% gap)

Berdasarkan Tabel 8, machine learning klasik (terutama *Naive Bayes*) lebih efektif untuk dataset berukuran sedang pada tugas deteksi aduan publik. Deep learning tidak memberikan peningkatan

performa yang signifikan, namun membutuhkan sumber daya komputasi yang jauh lebih besar.

4. KESIMPULAN

Penelitian ini membandingkan performa *machine learning* dan *deep learning* dalam klasifikasi teks aduan masyarakat berbahasa Indonesia pada media sosial menggunakan *Naive Bayes*, *Support Vector Machine* (SVM), *Convolutional Neural Network* (CNN), *Long Short-Term Memory* (LSTM), dan *Generative Adversarial Network* (GAN). Berdasarkan hasil pengujian pada dataset berukuran 5.000 data dengan dua kelas, *Naive Bayes* menunjukkan performa terbaik dibandingkan model lainnya, dengan *Accuracy* sebesar 77,50%, *Precision* 77,41%, *Recall* 77,50%, dan *F1-Score* 77,31%. CNN memperoleh performa yang relatif mendekati dengan *F1-Score* 77,09%, sedangkan SVM, LSTM, dan GAN memperoleh *F1-Score* masing-masing sebesar 74,67%, 73,66%, dan 41,98%.

Hasil tersebut menunjukkan bahwa pada dataset berukuran sedang, teks berbahasa Indonesia, dan domain aduan masyarakat yang digunakan dalam penelitian ini, model *machine learning* klasik, khususnya *Naive Bayes*, dapat memberikan performa yang lebih baik dibandingkan model *deep learning* yang diuji. Oleh karena itu, *Naive Bayes* dapat dipertimbangkan sebagai pendekatan yang efektif dan relatif ringan untuk skenario klasifikasi aduan pada karakteristik dataset penelitian ini, bukan sebagai rekomendasi yang berlaku secara umum untuk seluruh kasus klasifikasi teks. Performa GAN yang rendah menunjukkan bahwa penggunaan *Discriminator* sebagai classifier memerlukan pengembangan lebih lanjut, sedangkan hasil LSTM menunjukkan perlunya perhatian terhadap risiko *overfitting* pada dataset dengan jumlah data terbatas.

Penelitian selanjutnya dapat memperluas ukuran dan variasi dataset, menambahkan kategori aduan yang lebih spesifik, melakukan optimasi hiperparameter, serta mengevaluasi model berbasis *transformer* seperti BERT atau IndoBERT untuk mengetahui apakah pendekatan tersebut dapat meningkatkan performa klasifikasi pada domain aduan masyarakat berbahasa Indonesia.

REFERENSI

[1] M. Purba, F. Ilmu Komputer, and U. Sjakhyakirti, "Klasifikasi Dataset Teks Pengaduan Masyarakat Terhadap Pemerintah di Sosial Media Menggunakan Logistic Regression Article Info ABSTRAK," *JSAI: Journal Scientific and Applied Informatics*, vol. 7, no. 1, p. 5, doi: 10.36085.

[2] S. Anadas and R. R. Suryono, "Analisis Sentimen Publik Terhadap Program Lapor Mas Wapres Periode 2024 di Media Sosial X Menggunakan Metode Naive Bayes dan Support Vector Machine (SVM)," *Jurnal Pendidikan dan Teknologi Indonesia*, vol. 5, no. 8, pp. 2183–2192, Aug. 2025, doi: 10.52436/1.jpti.934.

[3] E. D. Madyatmadja, C. P. M. Sianipar, C. Wijaya, and D. J. M. Sembiring, "Classifying Crowdsourced Citizen Complaints through Data Mining: Accuracy Testing of k-Nearest Neighbors, Random Forest, Support Vector Machine, and AdaBoost," *Informatics*, vol. 10, no. 4, Dec. 2023, doi: 10.3390/informatics10040084.

[4] D. Rakhimzhanov, S. Belginova, and D. Yedilkhan, "Automated Classification of Public Transport Complaints via Text Mining Using LLMs and Embeddings," *Information (Switzerland)*, vol. 16, no. 8, Aug. 2025, doi: 10.3390/info16080644.

[5] F. Pradana Rachman, H. Santoso, and A. History, "Jurnal Teknologi dan Manajemen Informatika Perbandingan Model Deep Learning untuk Klasifikasi Sentiment Analysis dengan Teknik Natural Languange Processing Article Info ABSTRACT," vol. 7, no. 2, pp. 103–112, 2021, [Online]. Available: http://http://jurnal.unmer.ac.id/index.php/jtmi

[6] Purnomo Dandi, "KOMPARASI ALGORITMA RANDOM FOREST, NAÏVE BAYES, DAN SVM PADA SENTIMEN KEBIJAKAN PPN 12% COMPARISON OF RANDOM FOREST, NAÏVE BAYES, AND SVM ALGORITHMS IN SENTIMENT ANALYSIS OF THE 12% VAT POLICY," UNIVERSITAS SULAWESI BARAT, 2025.

[7] T. Chakraborty, U. Reddy K S, S. M. Naik, M. Panja, and B. Manvitha, "Ten years of generative adversarial nets (GANs): a survey of the state-of-the-art," Mar. 01, 2024, *Institute of Physics*. doi: 10.1088/2632-2153/ad1f77.

[8] Annas Rifai and Imelda, "Analysis Of Community Sentiment Towards The Implementation Of Zoning In New Student Admissions (PPDB) Using The Support Vector Machine (SVM) Algorithm," *Moneter: Jurnal Keuangan dan Perbankan*, no. Vol. 12 No. 1 (2024): APRIL, Apr. 2024, doi: 10.32832/moneter.v12i1.720.

[9] E. Utami and A. Dwi Hartanto, "ANALISIS SENTIMEN MASYARAKAT TERHADAP PELAKSANAAN P3K GURU DENGAN ALGORITMA NAIVE BAYES DAN DECISION TREE (THE ANALYSIS OF COMMUNITY SENTIMENT ON THE IMPLEMENTATION OF GOVERMENT EMPLOYEES WITH WORK AGREEMENT (P3K) TEACHERS WITH NAIVE BAYES ALGORITHM AND DECISION TREE)," *Jurnal TEKNIMEDIA : Teknologi Informasi dan Multimedia*, vol. Vol 22 No 1 (2026), Jun. 2022, doi: https://doi.org/10.25134/fon.v22i1.471.

[10] F. Ali, A. Setiawan, R. Nuraini, and S. Fathonah, "TINJAUAN SISTEMATIS LITERATUR: ANALISIS SENTIMEN TERHADAP PERSEPSI PUBLIK TENTANG KECERDASAN BUATAN DI MEDIA SOSIAL," *Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 5, Jul. 2025, doi: https://doi.org/10.36040/jati.v9i5.14963.

[11] R. Difandana and I. Imaduddin, "ANALISIS KOMPARATIF ALGORITMA NAIVE BAYES DAN SUPPORT VECTOR MACHINE DALAM KLASIFIKASI UJARAN KEBENCIAN DAN TEKS ABUSIF BERBAHASA INDONESIA," *Jurnal Pendidikan Bahasa dan Sastra Indonesia*, vol. Vol 22 No 1 (2026), Mar. 2026, doi: https://doi.org/10.25134/fon.v22i1.471.