Optimasi Ukuran Dataset untuk Analisis Sentimen Menggunakan Teknik Pembelajaran Mesin dan Pembelajaran Mendalam

St. Nur Halisah Duli[1], Fachrim Irhamna Rahman[2], Titin Wahyuni[3]
[1,2,3]Informatika, Universitas Muhammadiyah Makassar, Makassar, 90221, Indonesia
105841109020@student.unismuh.ac.id, Fachrim141020@unismuh.ac.id, titinwahyuni@unismuh.ac.id

Received: Agustus 01,2025; Accepted: Agustus 01, 2025; Published: 30 September, 2025

Abstract

This research aims to optimize the size of the dataset used in sentiment analysis through the application of machine learning and deep learning techniques. The machine learning methods used include Naive Bayes, Logistic Regression, and Support Vector Machine, while Convolutional Neural Network is used for the deep learning method. The data used in this research comes from a Google Maps review of several tourist attractions, such as Bugis Waterpark, Akcaus, Tanjung Bayang, Bosowa Beach, and Taman Wisata. The data pre-processing stage includes data cleaning, casefolding, stopword removal, tokenization, and stemming. Testing was carried out with nine different dataset sizes (4500, 4000, 3500, 3000, 2500, 2000, 1500, 1000, and 500) and divided the training data and test data in a ratio of 90:10, 80:20, and 70:30. The test results show that Logistic Regression with a dataset size of 1000 and a division of 90:10 achieves the highest level of accuracy, namely 85%. This study concludes that the optimal dataset size varies depending on the method used and underscores the importance of choosing the right dataset size to improve sentiment analysis performance.

Keywords: *Sentiment Analysis, Convolutional Neural Network, Naïve Bayes, Logistic Regression, Support Vector Machine.*

Abstrak

Penelitian ini bertujuan untuk mengoptimalkan ukuran dataset yang digunakan dalam analisis sentimen melalui penerapan teknik pembelajaran mesin dan pembelajaran mendalam. Metode pembelajaran mesin yang digunakan mencakup Naive Bayes, Regresi Logistik, dan Support Vector Machine, sedangkan Convolutional Neural Network digunakan untuk metode pembelajaran mendalam. Data yang digunakan dalam penelitian ini berasal dari ulasan Google Maps mengenai beberapa tempat wisata, seperti Bugis Waterpark, Akkarena, Tanjung Bayang, Pantai Bosowa, dan Wisata Kebun. Tahap pra-pemrosesan data melibatkan pembersihan data, case folding, penghapusan stopwords, tokenisasi, dan stemming. Pengujian dilakukan dengan sembilan ukuran dataset yang berbeda (4500, 4000, 3500, 3000, 2500, 2000, 1500, 1000, dan 500) serta pembagian data latih dan data uji dengan rasio 90:10, 80:20, dan 70:30. Hasil pengujian menunjukkan bahwa Regresi Logistik dengan ukuran dataset 1000 dan pembagian 90:10 mencapai tingkat akurasi tertinggi sebesar 85%. Studi ini menyimpulkan bahwa ukuran dataset yang optimal bervariasi tergantung pada metode yang digunakan dan menggarisbawahi pentingnya pemilihan ukuran dataset yang tepat untuk meningkatkan performa analisis sentimen.

Kata kunci*: Analisis sentimen, Convolutional Neural Network, Naïve Bayes, Regresi Logistik, Support Vector Machine.*

1. Pendahuluan

Analisis sentimen telah menjadi topik populer dalam Pemrosesan Bahasa Alami (NLP) karena efektivitasnya dalam mengungkap opini publik. Ini melibatkan penggunaan teknik NLP untuk menganalisis dan memahami opini, sikap, dan emosi yang diekspresikan dalam teks. Hal ini telah menyebabkan pertumbuhan pesat dalam penelitian dan aplikasi analisis sentiment [1].

Dalam proses pengambilan keputusan, informasi dapat dikategorikan sebagai fakta atau opini. Fakta adalah pernyataan objektif tentang kejadian, seringkali didukung oleh bukti, sedangkan opini adalah subyektif dan mencerminkan pandangan pribadi berdasarkan persepsi dan asumsi. [2].

Pembelajaran mesin untuk analisis sentimen melibatkan algoritma seperti Naive Bayes, Regresi Logistik, dan Support Vector Machine, yang belajar mendeteksi emosi dalam teks tanpa intervensi manusia. Pembelajaran mendalam, sebagai subbidang pembelajaran mesin, menggunakan jaringan saraf buatan untuk memodelkan pola data yang kompleks. Convolutional Neural Networks (CNN) adalah jenis algoritma pembelajaran mendalam yang digunakan dalam studi ini [4]. Dalam penelitian ini pada metode Deep Learning digunakan algoritma CNN.

Peneliti sering menghadapi kebingungan tentang kapan harus menggunakan pembelajaran mendalam dibandingkan pembelajaran mesin. Pembelajaran mendalam membutuhkan sumber daya komputasi yang signifikan, sehingga kurang cocok untuk dataset kecil. Studi ini membandingkan kinerja tiga algoritma pembelajaran mesin (Support Vector Machine, Naïve Bayes, dan Regresi Logistik) dan satu algoritma pembelajaran mendalam (Convolutional Neural Network) pada berbagai ukuran dataset.

2. Metode Penelitian

2.1 Analisis Sentimen

Dalam penambangan teks, pemrosesan bahasa alami, dan kecerdasan buatan, analisis sentimen memainkan peran yang sangat penting. Tujuan utama analisis sentimen adalah untuk mengenali, mengekstrak, dan memproses data teks secara otomatis guna menghasilkan informasi yang berguna. Disiplin ilmu ini juga berfokus pada analisis opini, sikap, evaluasi, dan penilaian terhadap peristiwa, topik, organisasi, dan individu [5].

Analisis sentimen adalah proses otomatis untuk mengekstraksi, memproses, dan memahami data teks yang tidak terstruktur untuk mendapatkan informasi sentimen yang terkandung dalam sebuah kalimat opini atau pendapat [6].

2.2 Machine Learning

Pembelajaran Mesin dilengkapi dengan serangkaian aturan program yang dijalankan oleh algoritma. Teknik ini dapat dikategorikan sebagai instruksi yang diproses dan dipelajari secara otomatis untuk menghasilkan keluaran yang optimal, semuanya dilakukan tanpa campur tangan manusia. Proses ini sepenuhnya otomatis, mengubah data menjadi beberapa pola yang kemudian dimasukkan ke dalam sistem untuk mendeteksi masalah secara otomatis. [7]. Pada penelitian ini, teknik Pembelajaran Mesin yang digunakan meliputi Naïve Bayes, Regresi Logistik, dan Support Vector Machine. Naïve Bayes bekerja dengan memprediksi kemungkinan terjadinya suatu peristiwa di masa depan berdasarkan data historis yang ada. [8]. Analisis regresi logistik adalah salah satu jenis analisis regresi di mana variabel respon bersifat kategorikal dan variabel prediktor bisa kategorikal atau numerik. Jika variabel respon terdiri dari dua kategori, analisis ini disebut regresi logistik biner. Dalam regresi logistik biner, variabel respon memiliki dua nilai: 1 untuk kejadian sukses dan 0 untuk kejadian gagal [9]. Prinsip kerja Support Vector Machine (SVM) adalah mencari ruang pemisah paling optimal untuk memisahkan data ke dalam kelas yang berbeda. Kinerja SVM sangat bergantung pada fungsi kernel dan parameter yang diterapkan [10]. Awalnya, algoritma ini hanya mampu melakukan klasifikasi biner, tetapi sekarang telah dikembangkan lebih lanjut sehingga dapat mengklasifikasikan banyak kelas sekaligus. Selain untuk klasifikasi, SVM juga bisa digunakan untuk regresi dan deteksi outlier [11].

2.3 Deep Learning

Deep learning berbeda dari metode machine learning konvensional karena secara otomatis mampu mewakili data seperti gambar, video, atau teks tanpa memerlukan penerapan aturan kode atau pengetahuan domain manusia [3].

Pada penelitian ini, digunakan Deep Learning dalam bentuk Convolutional Neural Network (CNN). CNN adalah jenis jaringan saraf tiruan yang menggunakan umpan balik untuk menjaga struktur hierarkisnya. CNN belajar tentang representasi fitur internal dan mengeneralisasi fitur-fitur dalam masalah gambar secara keseluruhan, seperti pengenalan objek dan tugas computer vision [12].

2.4 Klasifikasi

Klasifikasi melibatkan evaluasi objek data untuk menempatkannya ke dalam kelas tertentu dari serangkaian kelas yang ada. Proses klasifikasi dalam data mining terdiri dari beberapa tahap, antara lain [13]:

1. Pembangunan Model

Dalam tahap ini, sebuah model dibuat untuk menyelesaikan masalah klasifikasi dari kelas atau atribut dalam data. Ini merupakan fase pelatihan, di mana data latih dianalisis dengan menggunakan algoritma klasifikasi, sehingga model pembelajaran direpresentasikan dalam bentuk aturan klasifikasi.

2. Penerapan Model

Dalam tahap ini, model yang telah dibuat sebelumnya digunakan untuk menentukan atribut atau kelas dari data baru yang belum diketahui atribut atau kelasnya sebelumnya. Tahap ini bertujuan untuk memperkirakan keakuratan aturan klasifikasi terhadap data uji. Jika model dapat diterima, maka aturan tersebut dapat diaplikasikan untuk mengklasifikasikan data baru.

2.5 Supervised Learning

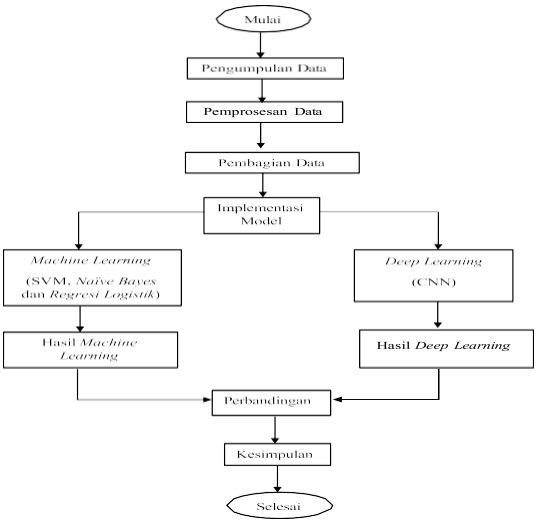
Supervised learning adalah metode dalam machine learning dan kecerdasan buatan yang memanfaatkan kumpulan data yang telah diberi label. Data-label ini membantu melatih algoritma untuk mengklasifikasikan data atau memprediksi hasil dengan akurasi yang tinggi. Data berlabel adalah data asli yang diperkaya dengan satu atau lebih informasi tambahan untuk memberikan konteks, sehingga mesin pembelajaran dapat mengacu pada informasi tersebut. Dengan menggunakan input dan output yang sudah diberi label, model mampu mengevaluasi tingkat keakuratannya dan terus memperbarui pengetahuannya dari waktu ke waktu [14].

2.6 Tensorflow

TensorFlow adalah kerangka kerja machine learning yang dirancang untuk bekerja dalam skala besar dan di lingkungan yang heterogen. TensorFlow digunakan untuk melakukan eksperimen dengan model deep learning, melatih model pada dataset yang sangat besar, dan mengoptimalkannya untuk produksi. Selain itu, TensorFlow juga mendukung pelatihan dan inferensi yang berskala besar dengan menggunakan ratusan server yang dilengkapi dengan Graphic Processing Unit (GPU) untuk pelatihan yang lebih efisien [15].

2.7 Perancangan Sistem

Flowchart atau diagram alur adalah gambaran grafis yang memvisualisasikan urutan langkah-langkah dan keputusan yang diperlukan untuk mengeksekusi suatu proses dalam sebuah program. Setiap langkah direpresentasikan dalam bentuk diagram dan dihubungkan oleh garis atau panah untuk menggambarkan alur proses.



Gambar 1. Proses Penelitian

2.8 Teknik Pengujian Sistem

Pengujian dilakukan menggunakan 9 set data berukuran beragam dan 4 metode yang berbeda, dengan penekanan pada akurasi sebagai metrik evaluasi utama. Akurasi adalah pengukuran seberapa baik model dapat mengklasifikasikan semua kelas sentimen (positif, negatif, dan netral) dengan tepat dalam set data. Semakin tinggi nilai akurasi, semakin baik kinerja model dalam melakukan klasifikasi.

$$Akurasi = \frac{Jumlah\ Prediksi\ yang\ benar}{Total\ Data\ Uji} \times 100\% \qquad (1)$$

2.9 Teknik Analisis Data

Analisis data adalah tahapan sistematis untuk menyusun dan menggali makna dari informasi yang diperoleh melalui wawancara, observasi, dan dokumentasi. Tujuan dari analisis data adalah memfasilitasi pemahaman informasi, baik bagi peneliti maupun pihak lain yang terlibat atau membaca hasil analisis tersebut.

3. Hasil dan Pembahasan

3.1 Pengambilan Data

Data ulasan diperoleh dengan menggunakan Instant Data Scraper yang mengekstraksi ulasan dari Google Maps untuk beberapa lokasi wisata, termasuk Pantai Akkarena, Tanjung Bayang, Pantai Bosowa, Wisata Kebun, dan Bugis Waterpark Adventure. Total ulasan yang berhasil dikumpulkan dari kelima lokasi wisata tersebut adalah 4500.

3.2 Pelabelan Data

Pelabelan sentimen dilakukan secara manual untuk mengidentifikasi pola dan karakteristik dalam teks yang menunjukkan sentimen positif, negatif, atau netral. Ulasan yang diperoleh dari Google Maps disimpan dalam atribut ulasan, sedangkan nilai klasifikasi seperti positif, negatif, atau netral disimpan dalam atribut label.

3.3 Preprocessing

a. Cleaning (Pembersihan)

Tahap pembersihan data melibatkan proses penghapusan tanda baca seperti koma, titik, tanda tanya, tanda seru, bintang, dan pagar dari teks atau data.

b. Transform Case

Transform Case adalah langkah dalam pemrosesan teks di mana karakter huruf yang awalnya dalam huruf besar (uppercase) diubah menjadi huruf kecil (lowercase).

c. Tokenizing

Tokenizing adalah langkah dalam pemrosesan teks yang melibatkan pembagian teks menjadi unit-unit yang lebih kecil yang disebut token. Unit-unit ini dapat berupa kata-kata, frasa, atau entitas lain.

d. Stopword

Stopword adalah langkah dalam pra-pemrosesan teks yang melibatkan eliminasi kata-kata umum dan sering muncul dalam teks, seperti kata depan, kata penghubung, dan kata-kata lain yang tidak memberikan kontribusi signifikan terhadap pemahaman makna teks.

e. Stemming

Saat melakukan stemming, bagian tambahan seperti akhiran, awalan, atau imbuhan pada kata-kata dihilangkan sehingga hanya bagian inti yang tersisa.

3.4 Perbandingan Metode

Setelah melakukan pengujian, diperoleh hasil akurasi dari metode Naive Bayes, regresi logistik, support vector machine, dan convolutional neural network dengan menggunakan sembilan ukuran data yang berbeda, yaitu 4500, 4000, 3500, 3000, 2500, 2000, 1500, 1000, dan 500, dengan pembagian data latih dan data uji sebesar 90:10, 80:20, dan 70:30. Dari hasil pengujian tersebut, kita dapat menentukan ukuran data yang tepat dalam penggunaan Machine Learning dan Deep Learning dalam melakukan analisis sentimen teks. Ukuran data 1000 dengan pembagian data 90:10 menunjukkan hasil yang paling optimal untuk analisis sentimen teks, dengan menggunakan metode regresi logistik yang mencapai akurasi tertinggi sebesar 85%.

4. Kesimpulan

Setelah pengujian dilakukan, hasil akurasi dari metode Naive Bayes, regresi logistik, support vector machine, dan convolutional neural network diperoleh dengan menggunakan sembilan ukuran data yang berbeda: 4500, 4000, 3500, 3000, 2500, 2000, 1500, 1000, dan 500. Data dipecah menjadi data latih dan data uji dengan perbandingan 90:10, 80:20, dan 70:30. Hasil pengujian ini memungkinkan kita untuk menentukan ukuran data yang optimal dalam penggunaan Machine Learning dan Deep Learning untuk analisis sentimen teks. Ukuran data 1000 dengan pembagian data 90:10 menunjukkan hasil yang paling baik untuk analisis sentimen teks, dengan menggunakan metode regresi logistik yang mencapai akurasi tertinggi sebesar 85%.

Referensi

[1] A. Nurzahputra And A. Muslim, *Analisis Sentimen Pada Opini Mahasiswa Menggunakan Natural Language Processing*. 2016.

[2] A. S. Stone And F. Fathoni, “Analisis Sentiment Pelanggan Terhadap Penilaian Produk Pada Toko Online Shop Amreta Menggunakan Metode Naïve Bayes Classification,” *Jurnal Media Informatika Budidarma*, Vol. 6, No. 3, P. 1590, Jul. 2022, Doi: 10.30865/Mib.V6i3.4436.

[3] A. Raup, W. Ridwan, Y. Khoeriyah, Q. Yuliati Zaqiah, And U. Islam Negeri Sunan Gunung Djati Bandung, “Deep Learning Dan Penerapannya Dalam Pembelajaran,” 2022. [Online]. Available: Http://Jiip.Stkipyapisdompu.Ac.Id

[4] F. Fitroh And F. Hudaya, “Systematic Literature Review: Analisis Sentimen Berbasis Deep Learning,” *Jurnal Nasional Teknologi Dan Sistem Informasi*, Vol. 9, No. 2, Pp. 132–140, Aug. 2023, Doi: 10.25077/Teknosi.V9i2.2023.132-140.

[5] F. Bei And S. Saepudin, “Analisis Sentimen Aplikasi Tiket Online Di Play Store Menggunakan Metode Support Vector Machine (Svm),” 2021.

[6] P. Arsi And R. Waluyo, “Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (Svm),” Vol. 8, No. 1, Pp. 147–156, 2021, Doi: 10.25126/Jtiik.202183944.

[7] Rahmadini, E. Lubis Lorencis Erika, A. Priansyah, Y. R.W.M, And T. Meutia, “Penerapan Data Mining Untuk Memprediksi Harga Bahan Pangan Di Indonesia Menggunakan Algoritma K-Nearest Neighbor,” 2023.

[8] A. Herdhianto, “Sentiment Analysis Menggunakan Naïve Bayes Classifier (Nbc) Pada Tweet Tentang Zakat,” 2020.

[9] D. A. Ayungtyas, “Klasifikasi Menggunakan Metode Regresi Logistik Dan Support Vector Machine,” 2017.

[10] V. Seplifriskila Tampubolon, E. J. Sagala, P. Manajemen, B. Telekomunikasi, And D. Informatika, "Pengaruh Kepuasan Kerja Dan Komitmen Organisasi Terhadap Turnover Intention Pada Karyawan Pt. Bum Divisi Pmks," *Jurnal Business Management Journal*, Vol. 16, No. 2, Pp. 65–80, 2020, Doi: 10.30813/Bmj.

[11] F. Abdusyukur, "Penerapan Algoritma Support Vector Machine (Svm) Untuk Klasifikasi Pencemaran Nama Baik Di Media Sosial Twitter," *Komputa : Jurnal Ilmiah Komputer Dan Informatika*, Vol. 12, No. 1, 2023.

[12] D. Tri Hermanto, A. Setyanto, And E. T. Luthfi, "Algoritma Lstm-Cnn Untuk Sentimen Klasifikasi Dengan Word2vec Pada Media Online Lstm-Cnn Algorithm For Sentiment Clasification With Word2vec On Online Media," 2021.

[13] Mustika *Et Al.*, *Data Mining Dan Aplikasinya*. 2021.

[14] K. Kristiawan, D. D. Somali, T. A. Linggan Jaya, And A. Widjaja, "Deteksi Buah Menggunakan Supervised Learning Dan Ekstraksi Fitur Untuk Pemeriksa Harga," *Jurnal Teknik Informatika Dan Sistem Informasi*, Vol. 6, No. 3, Dec. 2020, Doi: 10.28932/Jutisi.V6i3.3029.

[15] M. Irfan, "Penerapan Chatbot Untuk Mendeteksi Kemiripan Gambar Asli Atau Manipulasi Menggunakan Metode Metadata Dan Error Level Analysis (Ela) Sebagai Komputasi Cnn," 2020.