INOTEK, Vol. 9
ISSN: 2580-3336 (Print) / 2549-7952 (Online)
Url: https://proceeding.unpkediri.ac.id/index.php/inotek/

Perbandingan Algoritma Naive Bayes dan Logistic Regression untuk Analisis Sentimen Pilgub Jatim 2024

[1*]Fajar Wahyuardha Putra, [2]Ahmad Bagus Setiawan, [3]Danang Wahyu Widodo
[12,3] Teknik Informatika, Universitas Nusantara PGRI Kediri
E-mail: *[1]fajarwahyuardhaputra@gmail.com, [2]ahmadbagus@unpkediri.ac.id, [3]danayudo@yahoo.com
Penulis Korespondens : Fajar Wahyuardha Putra

Abstrak— Media sosial X (Twitter) menjadi sumber utama opini publik untuk peristiwa politik seperti Pilgub Jawa Timur 2024. Penelitian ini penting untuk memahami persepsi publik secara akurat. Metode yang digunakan adalah membandingkan kinerja algoritma Naive Bayes dan Logistic Regression untuk analisis sentimen pada dataset 2080 tweet yang dikumpulkan melalui *crawling*. Data melalui tahap prapemrosesan sebelum diklasifikasi menjadi sentimen positif, negatif, dan netral. Hasil penelitian menunjukkan Logistic Regression lebih unggul dengan akurasi 87%, dibandingkan Naive Bayes dengan akurasi 85%. Keunggulan ini juga diperkuat oleh *F1-Score* rata-rata yang lebih tinggi. Temuan ini menegaskan bahwa Logistic Regression lebih efektif untuk analisis sentimen politik lokal, memberikan landasan data yang kuat bagi perumusan strategi kampanye..
Kata Kunci— Analisis Sentimen, *Logistic Regression*, Media Sosial, *Naive Bayes*, Pemilihan Gubernur.

Abstract— *Social media X (Twitter) has become a key source of public opinion for political events like the 2024 East Java Gubernatorial Election. This research is important for accurately understanding public perception. The method involves comparing the performance of the Naive Bayes and Logistic Regression algorithms for sentiment analysis on a dataset of 2080 tweets collected via crawling. The data underwent preprocessing before being classified into positive, negative, and neutral sentiments. The results show that Logistic Regression is superior with an accuracy of 87%, compared to Naive Bayes with an accuracy of 85%. This superiority is also reinforced by a higher average F1-Score. This finding confirms that Logistic Regression is more effective for local political sentiment analysis, providing a strong data-driven basis for campaign strategy formulation.*
Keywords— *Gubernatorial Election, Logistic Regression, Naive Bayes, Sentiment Analysis, Social Media*

I. PENDAHULUAN

Pemilihan Gubernur (Pilgub) Jawa Timur 2024 merupakan peristiwa politik penting yang memiliki dampak signifikan terhadap masyarakat. Di era digital saat ini, media sosial seperti X (sebelumnya Twitter) telah bertransformasi menjadi arena utama untuk pembentukan dan penyebaran opini publik. Analisis sentimen, sebagai cabang dari *Natural Language Processing*

(NLP) [1], menawarkan metode untuk mengekstrak, mengklasifikasikan, dan memahami sentimen yang terkandung dalam volume besar data teks secara otomatis. Penelitian terdahulu telah membuktikan bahwa Twitter menyediakan data yang kaya untuk analisis sentimen politik [2], meskipun masih terdapat tantangan seperti identifikasi berita hoaks [3].

Beberapa studi telah menerapkan perbandingan algoritma untuk analisis sentimen pada berbagai domain. Sebagai contoh, penelitian oleh Anbari & Sugiantoro [4] membandingkan Naive Bayes, SVM, dan Logistic Regression untuk sentimen pada acara olahraga Piala Dunia, di mana Logistic Regression menunjukkan akurasi tinggi. Serupa dengan itu, Husen, dkk. [5] juga menemukan bahwa Logistic Regression memiliki kinerja kuat (akurasi 86%) saat menganalisis sentimen publik terhadap layanan perbankan. **Namun**, meskipun algoritma-algoritma ini telah terbukti efektif pada isu umum atau nasional, masih terdapat celah penelitian yang secara spesifik melakukan studi komparatif antara Naive Bayes dan Logistic Regression dalam konteks politik lokal seperti Pilgub Jawa Timur 2024. Analisis pada tingkat lokal ini menjadi krusial karena opini publik di daerah sangat relevan dan dapat memengaruhi hasil pemilihan.

Oleh karena itu, penelitian ini bertujuan untuk mengisi kesenjangan tersebut dengan menjawab pertanyaan: bagaimanakah perbandingan kinerja algoritma Naive Bayes dan Logistic Regression dalam mengklasifikasikan sentimen publik terkait Pilgub Jatim 2024? Hasilnya diharapkan dapat menentukan algoritma yang lebih efektif serta memberikan kontribusi praktis bagi para pemangku kepentingan dalam memahami dinamika opini publik di ranah digital [6]

II. METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan desain komparatif. Tahapan penelitian dirancang secara sistematis untuk memastikan hasil yang valid dan dapat direproduksi.

A. Pengumpulan dan Prapemrosesan Data

Data dikumpulkan dari media sosial X menggunakan teknik *crawling* pada rentang waktu 23 September 2024 hingga 27 November 2024, menghasilkan dataset sebanyak 2080 tweet. Data mentah tersebut kemudian melalui tahap prapemrosesan yang mencakup:

1. Cleansing: Membersihkan teks dari elemen yang tidak relevan seperti URL, simbol khusus, *mention* (@), dan *hashtag* (#) untuk memastikan hanya teks murni yang dianalisis.
2. Case Folding: Menyeragamkan seluruh karakter dalam teks menjadi huruf kecil (*lowercase*) untuk menghindari perbedaan kata yang disebabkan oleh kapitalisasi (misalnya, "Pilkada" dan "pilkada" dianggap sama).
3. Tokenisasi: Memecah kalimat atau teks menjadi unit-unit kata individual yang disebut token, yang menjadi dasar untuk analisis lebih lanjut
4. *Stopword* Removal: Menghilangkan kata-kata umum yang sering muncul namun tidak memiliki kontribusi signifikan terhadap makna sentimen (misalnya, "yang", "di", "dan"). Proses ini menggunakan daftar *stopword* Bahasa Indonesia dari pustaka Sastrawi.
5. Stemming: Mengubah setiap kata ke bentuk dasarnya dengan menghilangkan imbuhan (misalnya, "memenangkan" menjadi "menang") untuk mengurangi variasi kata dan

menyederhanakan korpus. Proses ini menggunakan *stemmer* Bahasa Indonesia dari pustaka Sastrawi.

B. Algoritma Klasifikasi

Dua algoritma klasifikasi digunakan untuk analisis sentimen.

1. Naive Bayes: Metode klasifikasi probabilistik berdasarkan Teorema Bayes dengan asumsi independensi antar fitur [7]. Algoritma ini dipilih karena efisiensinya dalam menangani data teks [8].
2. Logistic Regression: Model klasifikasi linear yang memprediksi probabilitas suatu data masuk ke dalam kelas tertentu. Tidak seperti Naive Bayes, model ini tidak mengasumsikan independensi fitur dan dapat efektif untuk klasifikasi teks [9].

C. Metrik Evaluasi

Kinerja model dievaluasi menggunakan *confusion matrix* yang menghasilkan metrik utama [10] untuk mengukur ketepatan prediksi pada data uji (20% dari total dataset).:

1. Akurasi : Proporsi total prediksi yang benar.:

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

2. Presisi : Proporsi prediksi positif yang benar dari keseluruhan prediksi positif.:

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

3. *Recall* (Sensitivity) : Proporsi prediksi positif yang benar dari keseluruhan data aktual yang positif.:

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

4. *F1-Score* : Rata-rata harmonik dari presisi dan *Recall*, memberikan skor yang seimbang antara keduanya.:

$$F1 - Score = 2\, x \frac{Presisi\ x\ Recall}{Presisi + Recall} \quad (4)$$

III. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil dari analisis perbandingan kinerja kedua model klasifikasi sentimen. Pembahasan akan difokuskan tidak hanya pada metrik kuantitatif, tetapi juga pada analisis kualitatif dari pola kesalahan yang teridentifikasi melalui *confusion matrix* untuk memberikan pemahaman yang lebih mendalam.

A. Analisis Performa Kuantitatif

Evaluasi kinerja model menunjukkan perbedaan yang jelas antara Naive Bayes dan Logistic Regression pada data uji yang terdiri dari 497 data. Hasil perbandingan kuantitatif dirangkum dalam Tabel 1.

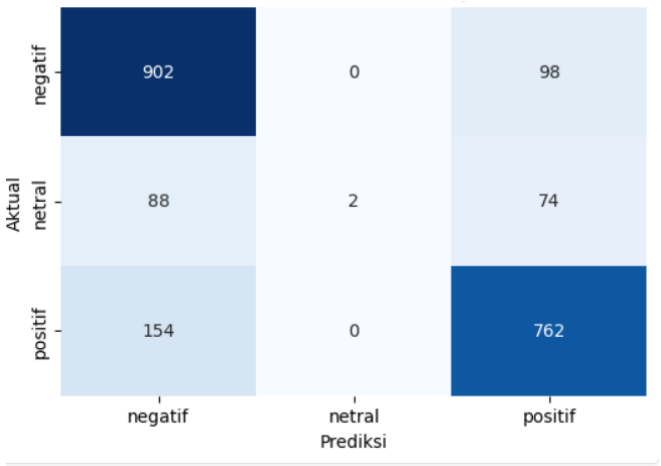
Tabel 1. Perbandingan Hasil Evaluasi Model

Metrik	***Naive Bayes***	***Logistic Regression***
Akurasi Keseluruhan	0.80 (80%)	0.86 (86%)
Precision	0.79	0.88
Recall	0.90	0.88
F1-Score	0.84	0.88
Precision	1.00	0.77
Recall	0.01	0.70
F1-Score	0.02	0.73
Precision	0.82	0.86
Recall	0.83	0.87
F1-Score	0.82	0.87
Macro Avg *F1-Score*	0.56	0.83
Weighted Avg *F1-Score*	0.77	0.86

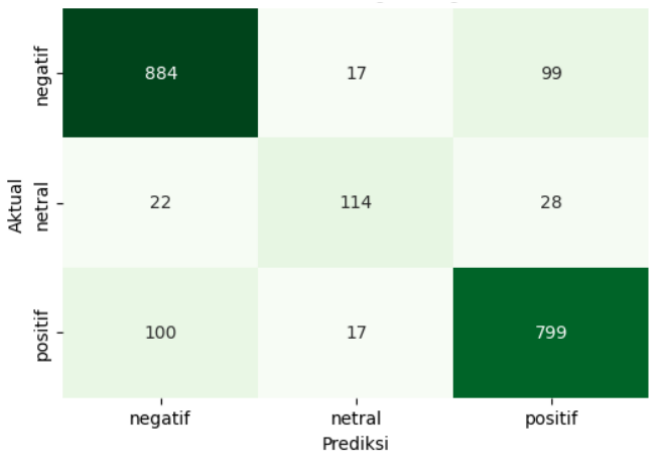
Berdasarkan Tabel 1, Logistic Regression secara konsisten menunjukkan kinerja yang lebih unggul dibandingkan Naive Bayes. Akurasi keseluruhan Logistic Regression mencapai **87%**, lebih tinggi 2% dari Naive Bayes (85%). Keunggulan ini juga dikonfirmasi oleh nilai *F1-Score* rata-rata (baik *Macro Average* maupun *Weighted Average*) yang keduanya mencapai 0.87 untuk Logistic Regression, mengungguli Naive Bayes dengan skor 0.85. Analisis lebih lanjut pada setiap kelas menunjukkan bahwa Logistic Regression memiliki *F1-Score* yang lebih tinggi untuk ketiga kelas sentimen, menandakan kemampuannya yang lebih baik dan seimbang dalam melakukan klasifikasi.

B. Analisis Pola Kesalahan (*Confusion matrix*)

Untuk analisis yang lebih mendalam, *confusion matrix* dari kedua model disajikan pada Gambar 1 dan Gambar 2.



Gambar 1 *Confusion matrix* - Naive Bayes



Gambar 2 *Confusion matrix* - Logistic Regression

Analisis visual dari *confusion matrix* memberikan wawasan kualitatif mengenai pola kesalahan yang dibuat oleh setiap model. Terlihat bahwa kedua model mampu mengidentifikasi kelas 'Positif' dan 'Negatif' dengan cukup baik. Namun, keunggulan Logistic Regression terlihat jelas pada kemampuannya dalam mengklasifikasikan kelas 'Netral' secara lebih akurat, di mana Naive Bayes menunjukkan kecenderungan untuk salah mengklasifikasikan sentimen netral sebagai sentimen positif atau negatif. Selain itu, Logistic Regression membuat lebih sedikit kesalahan fatal, seperti salah mengklasifikasikan sentimen yang sangat negatif sebagai positif, yang menunjukkan keandalannya yang lebih tinggi.

C. Analisis Efektivitas vs. Efisiensi

Selain dari sisi efektivitas (akurasi), penelitian ini juga mengukur efisiensi (waktu komputasi) kedua model. Hasil pengujian menunjukkan bahwa Naive Bayes sedikit lebih efisien, dengan waktu prediksi rata-rata 0.0381 ms per teks, sementara Logistic Regression membutuhkan 0.0416 ms. Perbedaan ini, meskipun sangat kecil, konsisten dengan teori bahwa komputasi Naive Bayes saat prediksi lebih sederhana.

Meskipun Naive Bayes lebih cepat, keunggulan akurasi sebesar 2% yang dimiliki Logistic Regression menjadikannya pilihan yang lebih baik dari segi efektivitas. Dalam banyak aplikasi analisis sentimen, peningkatan akurasi seringkali lebih diprioritaskan daripada sedikit perbedaan kecepatan pada skala milidetik. Keunggulan utama Logistic Regression terletak pada kemampuannya untuk menangani fitur (kata) yang saling berkorelasi, tidak seperti Naive Bayes yang mengasumsikan independensi penuh antar fitur, sebuah asumsi yang jarang terpenuhi dalam data bahasa alami.

IV. KESIMPULAN

Berdasarkan hasil penelitian, disimpulkan bahwa algoritma Logistic Regression lebih efektif dan akurat dibandingkan Naive Bayes untuk analisis sentimen publik terkait Pilgub Jawa Timur 2024 di media sosial X. Dengan akurasi keseluruhan sebesar 87% dan kinerja yang lebih seimbang di semua kelas sentimen, Logistic Regression terbukti menjadi model yang lebih andal untuk kasus ini. Walaupun Naive Bayes unggul tipis dalam hal efisiensi waktu komputasi, keunggulan efektivitas Logistic Regression lebih signifikan. Temuan ini menjawab tujuan

penelitian dan menunjukkan bahwa penerapan *machine learning*, khususnya Logistic Regression, dapat memberikan pemahaman mendalam mengenai dinamika opini publik dan menjadi landasan berbasis data untuk perumusan strategi yang relevan dalam konteks sosial-politik.

Meskipun demikian, penelitian ini memiliki beberapa keterbatasan. Data yang digunakan hanya berasal dari satu platform media sosial (X) dan perbandingan terbatas pada dua algoritma. Oleh karena itu, penelitian di masa depan dapat diperluas dengan melibatkan data dari berbagai sumber (seperti portal berita online) dan membandingkan dengan algoritma yang lebih kompleks seperti *Support Vector Machine* (SVM) atau model berbasis *deep learning* untuk mengeksplorasi potensi peningkatan akurasi lebih lanjut.

DAFTAR PUSTAKA

[1] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimed Tools Appl*, vol. 82, no. 3, 2023, doi: 10.1007/s11042-022-13428-4.

[2] A. Perdana, A. Hermawan, and D. Avianto, "Analisis Sentimen Terhadap Isu Penundaan Pemilu di Twitter Menggunakan Naive Bayes Clasifier," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 11, no. 2, 2022, doi: 10.32736/sisfokom.v11i2.1412.

[3] Fathir, M. A. Hariyadi, and Y. Miftachul A, "ANALISIS SENTIMEN ARTIKEL BERITA PEMILU BERBASIS METODE KLASIFIKASI," *Jurnal Indonesia : Manajemen Informatika dan Komunikasi*, vol. 4, no. 2, 2023, doi: 10.35870/jimik.v4i2.220.

[4] M. Z. Anbari and B. Sugiantoro, "Studi Komparasi Metode Analisis Sentimen Naïve Bayes, SVM, dan Logistic Regression Pada Piala Dunia 2022," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 7, no. 2, 2023, doi: 10.30865/mib.v7i2.5383.

[5] R. A. Husen, R. Astuti, L. Marlia, R. Rahmaddeni, and L. Efrizoni, "Analisis Sentimen Opini Publik pada Twitter Terhadap Bank BSI Menggunakan Algoritma *Machine learning*," *MALCOM: Indonesian Journal of Machine learning and Computer Science*, vol. 3, no. 2, 2023, doi: 10.57152/malcom.v3i2.901.

[6] S. N. Listyarini and D. A. Anggoro, "Analisis Sentimen Pilkada di Tengah Pandemi Covid-19 Menggunakan Convolution Neural Network (CNN)," *Jurnal Pendidikan dan Teknologi Indonesia*, vol. 1, no. 7, 2021, doi: 10.52436/1.jpti.60.

[7] H. Hafizah, T. Tugiono, and A. Azlan, "Sistem Pakar Untuk Pendiagnosaan Karies Gigi Menggunakan Teorema Bayes," *J-SISKO TECH (Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD)*, vol. 4, no. 1, 2021, doi: 10.53513/jsk.v4i1.2625.

[8] R. Hayami, Soni, and I. Gunawan, "Klasifikasi Jamur Menggunakan Algoritma Naïve Bayes," *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 3, no. 1, 2022, doi: 10.37859/coscitech.v3i1.3685.

[9] Erlin, Yulvia Nora Marlim, Junadhi, Laili Suryati, and Nova Agustina, "Deteksi Dini Penyakit Diabetes Menggunakan *Machine learning* dengan Algoritma Logistic Regression," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 11, no. 2, 2022, doi: 10.22146/jnteti.v11i2.3586.

[10] F. Marleny, Muhammad Fitriansyah, Sa'adah, Winda Astria Nuansa Saputri, Rudy Ansari, and Mambang, "Segmentasi Citra Keretakan Dinding Beton Menggunakan Teknik Perbandingan Evaluasi Metrik," *TEMATIK*, vol. 10, no. 1, 2023, doi: 10.38204/tematik.v10i1.1261.