JURNAL INFORMATIKA DAN TEKNOLOGI KOMPUTER
Halaman Jurnal: https://journal.amikveteran.ac.id/index.php/jitek
Halaman UTAMA Jurnal : https://journal.amikveteran.ac.id/index.php

KOMPARASI ALGORITMA KLASIFIKASI NAIVE BAYES DAN K-NEAREST NEIGHBORS DALAM ANALISIS SENTIMEN TERHADAP OPINI FILM PADA TWITTER

Muhammad Muharrom
Fakultas Teknik dan Informatika / Program Studi Teknologi Informasi, muhammad.muu@bsi.ac.id,
Universitas Bina Sarana Informatika

ABSTRACT

The fact that social media is so unreliable does not prevent Twitter users from using the service. Twitter is one of a few social media platforms that allows users to engage in conversation, share information, or even reveal their true identities, such as when discussing a movie's plot. A tweet or comment about a movie that is posted on Twitter may be viewed as a tool to improve the quality of movie production. To understand this, one can use sentimen analysis to categorize as either negative or positive by comparing the Naive Bayes and k-Nearest Neighbors algorithms to determine which one is the most accurate. The results of the two algorithms' comparative testing reveal that the Nave Bayes algorithm has a higher rata-rata accuracy of 99.63% with an AUC of around 1.000, while the K-NN algorithm has a higher rata-rata accuracy of 99.25% with an AUC of 1.000.

Keywords: *Twitter, Tweets, Sentiment analysis, Data mining, Comparisons, naivebayes, K-Nearest Neighbors.*

Abstrak

Fakta bahwa media sosial sangat tidak dapat dihindarkan tidak menghalangi pengguna Twitter untuk menggunakan layanan ini. Twitter adalah salah satu dari sedikit platform media sosial yang memungkinkan pengguna untuk terlibat dalam percakapan, berbagi informasi, atau bahkan mengungkapkan identitas aslinya, seperti saat mendiskusikan plot film. Tweet atau komentar tentang film yang diposting di Twitter dapat dilihat sebagai alat untuk meningkatkan kualitas produksi film. Untuk memahami hal ini, seseorang dapat menggunakan analisis sentimen untuk mengkategorikan data sebagai negatif atau positif dengan membandingkan algoritma Naive Bayes dan k-Nearest Neighbors untuk menentukan mana yang paling akurat. Hasil pengujian komparatif kedua algoritma menunjukkan bahwa algoritma Nave Bayes memiliki akurasi rata-rata yang lebih tinggi yaitu 99,63% dengan AUC sekitar 1.000, sedangkan algoritme K-NN memiliki akurasi rata-rata yang lebih tinggi yaitu 99,25% dengan AUC sebesar 1.000.

Kata Kunci: Twitter, Tweet, Analisis sentimen, Data mining, Komparasi, naivebayes, K-Nearest Neighbors.

1. PENDAHULUAN

Pengguna dapat mengomentari satu sama lain secara *real time* di Twitter, sebuah situs jejaring sosial. Biasanya, 140 karakter "*Tweet*" atau "Kicauans", yang merupakan komunikasi singkat, digunakan untuk mendeskripsikan informasi di Twitter. [1] Pemanfaatan analisis sentimen dapat diterapkan pada postingan *Twitter* tentang film berbahasa Indonesia. Seringkali, akan ada *tweet* yang ditulis dengan buruk saat menulis opini tentang sebuah film. Berdasarkan beberapa faktor, antara lain penggunaan tulisan yang disingkat, bahasa kekinian atau bahasa gaul, kurangnya keterampilan menulis surat, dan penulisan opini yang kurang tepat. Beberapa faktor ini menunjukkan bahwa representasi kata harus dilakukan sebelum jajak pendapat. Frasa tersebut di atas akan diklasifikasikan sehingga dapat ditentukan apakah akan berkonotasi negatif atau positif. Sehingga menganalisis komentar atau pendapat. Dengan metode klasifikasi.

Pada penelitian sebelumnya, metode Naive Bayes digunakan untuk memunculkan kelas positif dan negatif pada komentar pengguna aplikasi Tokopedia di Playstore. Pengujian didasarkan pada skor kategori negatif, kategori positif, recall dan presisi analisis sentimen. dengan nilai kinerja presisi yang baik sebesar 97,13% dan nilai presisi sebesar 1. Class recall memberikan nilai sebesar 95,49% (kelas positif: negatif). Dan nilai AUC adalah 0,980[2].

Dari beberapa penelitian sebelumnya diatas terkait dengan analisa sentimen maka penulis menganalisa komparasi terkait dengan klasifikasi naive bayes dan k-nearest neighbors dan melakukan ujicoba mengenai akurasi tersebut, sehingga penulis mampu mendapatkan pengetahuan dari dua penggunaan algoritma

Received Februari 21, 2023; Revised Maret 20, 2023; Accepted Maret 27, 2023

keduanya yang menghasilkan nilai akurasi terbaik dalam menganalis sentimen opini pada *twitter* terhadap penilaian film.

2. TINJAUAN PUSTAKA

2.1. Hubungan antara twitter dengan penilaian terhadap film

Peningkatan pengguna Twitter meningkatkan jumlah tweet yang dikirim. Tweet ini dapat berisi opini dan komentar publik terkait bisnis, perilaku sosial, fenomena alam, bisnis, pendidikan, hiburan, dan bidang lainnya. Hal-hal yang berhubungan dengan hiburan adalah film-film yang sedang beredar. Pengguna memposting komentar dan pendapat tentang film populer saat ini melalui Twitter. Pengguna mengirimkan peringkat untuk film yang ditonton. Informasi berupa permintaan pengguna menjadi acuan bagi pengguna Twitter lainnya ketika ingin menonton film yang sama. Kicauan pengguna juga bisa menjadi rating film-film yang diproduksi oleh rumah produksi film. Selain itu, tweet yang disusun secara acak membuat sulit untuk menemukan opini positif, negatif, atau netral. [3].

2.2. Data Mining

Teknik menemukan koneksi dalam data yang tidak disadari konsumen dikenal sebagai data mining. Proses ini sebanding dengan pencarian di gundukan pasir terkait dengan emas. Data mining juga sebagai proses mencari informasi menarik yang tersembunyi yang bertujuan untuk mengungkap hukum, korelasi, atau pola lain dalam kumpulan data yang sangat besar.[4], berikut tahapan yang dapat dilakukan:

1) *Data Selection*, pertama berupa *database* yang dapat digunakan untuk analisis data yang telah disimpan tetapi terkadang tidak dapat digunakan dikarena perbedaan struktur atau nilai.
2) Data Preprocessing adalah teknik pengelolaan data yang melibatkan pemilihan data terlebih dahulu (data cleaning). Pembersihan data seringkali melibatkan penggabungan data duplikat, menghilangkan data yang tidak penting, menemukan data yang hilang, dan memperbaiki kesalahan data seperti salah ketik. Dengan menghapus data yang tidak berguna, algoritma penambangan data dapat diperkuat dan ditingkatkan.
3) Transformasi data dimasukkan atau disimpan dalam format yang sesuai untuk penambangan data dan berfungsi dengan aplikasi yang digunakan untuk penambangan data. Sebelum dapat digunakan, beberapa teknik penambangan data membutuhkan format data yang bersih dari bias dalam pengumpulan data.
4) Data Mining adalah praktik menggunakan teknik dan algoritma tertentu untuk mencari pola atau informasi yang belum ditemukan dalam kumpulan data sebelumnya.
5) *Interpretasion /Evaluasi* yaitu Mengubah titik-titik atau pola yang diamati menjadi informasi yang lebih mudah dipahami oleh organisasi terkait.

2.3. Confusion Matrix

Merupakan proses menghitung keakuratan. Akurasi merupakan prediksi positif dari sebuah laporan kasus, dan juga dikatakan aktual jika benar – benar positif. *Accuracy* merupakan proses menghitung secara keseluruhan data untuk menghasilkan rasio prediksi benar (positif dan negatif).[5]

2.4. Kurva ROC

adalah suatu proses algoritma klasifikasi tertentu divisualisasikan berdasrkan kinerja dalam bentuk graf. Graf tersebut memberikan informasi tentang hubungan antara true positive dan false positive untuk menunjukkan threshold atau ambang batas sebagai proses penentuan proporsi data positif dan negatif.[6] Dalam kurva ROC, nilai FPR diwakili oleh sumbu x, sedangkan TPR diwakili sumbu y. Titik (0,0) pada kurva ROC menunjukkan bahwa keseimbangna antara kelas positif dan negatif, sedangkan titik (1,1) menunjukkan bahwa model memiliki performa sempurna dalam hubungan antara kelas positif dan negatif.

3. METODOLOGI PENELITIAN

Pada penelitian ini digunakan untuk menyelesaikan beberapa langkah dalam proses penelitian, yaitu:

3.1 *Data resource*

Proses penelitian yang ada untuk Analisis Sentimen Terhadap Opini Film Pada *Twitter* namun, algoritma yang paling akurat belum diketahui. Oleh karena itu, mencari nilai akurasi dari Naive Bayes dan k-NN untuk dapat menentukan relevansi opini di Twitter mana yang lebih akurat dalam perbandingannya. Pada data yang digunakan terdiri dari 200 *record* yang memiliki 100 data opini positif dan 100 data negatif. Berikut beberapa record opini film pada tweet:

Tabel 1. Dataset opini film pada twitter

Id	Sentiment	Text Tweet
1	Negative	Jelek filmnya... apalagi si ernest gak mutu bgt actingnya... film sampah
2	Negative	Film king Arthur ini film paling jelek dari seluruh cerita King Arthur
3	Negative	@beexkuanlin Sepanjang film gwa berkata kasar terus pada bapaknya
4	Negative	Ane ga suka fast and furious..menurutku kok jelek ya tu film
5	negative	@baekhyun36 kan gua ga tau film nya, lu bilang perang perangan/? Perang"an disebut ama rp yaoi jadi ambigu :v
...	.....	.....
196	positive	Fargo juga adaptasi dari film yang cukup berhasil. Season 1-nya the best! https://t.co/tkEEK3Evs9
197	positive	637.000 waw ini sangat keren flm horor dng jumlah penonton segini dlm waktu 4 hari @prillybie @danurmovie
198	positive	@filmziarah film yang tenang dan menghanyutkan. Salut dengan Mbah Ponco yg bisa membawakan karakter Mbah Sri dengan sangat baik. #FilmZiarah
199	positive	Film yg amat menarik. Kisah cinta & kesetiaan yg disajikan secara tidak biasa. Bikin kgn nenek. Recomended movie @filmziarah #meiberziarah
200	positive	Nntn @filmziarah , film bagus, ada kali 5 menit penonton gak beranjak tetep ngliatin layar pdhl film dah selesai, msh spicles sm endingnya

3.2 Pengolahan Awal Data

Saat ini, sebagian besar data berupa nilai numerik, yang terus menerus diubah menjadi nilai kategorikal dan dibangun menurut skala atau rentang untuk menghasilkan wilayah yang lebih kecil yang digunakan sebagai bahan pelatihan untuk algoritma naive bayes dan k-NN. menjadi . Menggunakan Rapid Miner memudahkan untuk mengklasifikasikan kumpulan data yang sudah tersedia.

3.3 Implementasi Naive Bayes dan K-NN

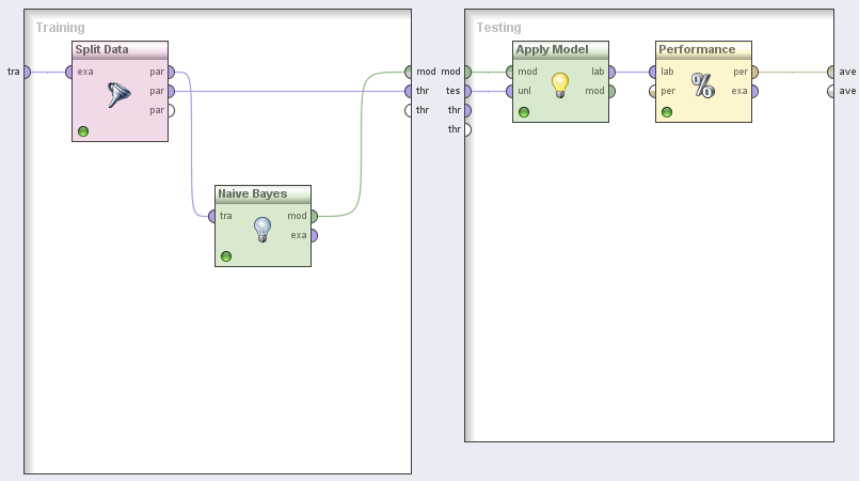
Pada fase ini, aplikasi yang tepat dipilih dan diterapkan untuk mencapai hasil terbaik. dan pemilihan algoritma yang tepat adalah algoritma Naive Bayes dan k-Nearest Neighbors. Langkah ini dilakukan dengan aplikasi Rapid Miner.

3.4 Analisis perbandingan hasil

Pada proses ini, hasil nilai algoritma dibandingkan untuk mengidentifikasi algoritma dengan akurasi tertinggi.

4. HASIL DAN PEMBAHASAN

Penelitian memiliki tujuan ini adalah untuk mempelajari lebih lanjut tentang algoritma yang memiliki Analisis Sentimen Opini Bioskop setinggi mungkin di Twitter. Info yang dianalisis adalah data yang dihasilkan dari umpan balik pengguna. Data yang dimaksud diperiksa menggunakan aplikasi Rapid Miner menggunakan Naive Bayes dan k-NN. kemudian membandingankan tingkat kinerja yang dihasilkan. Berdasarkan nilai Akurasi dan AUC. menggunakan operator split data, pada tools Rapid Miner dengan data training dan data testing dengan desain modelnya sebagai berikut:



Gambar 1. Model Pengujian Validasi Naive Bayes

Penelitian ini menggunakan 4 uji coba untuk metode algoritma Naive Bayes, berikut diantara satu contoh proses perhitungan menggunakan Rapid Miner.

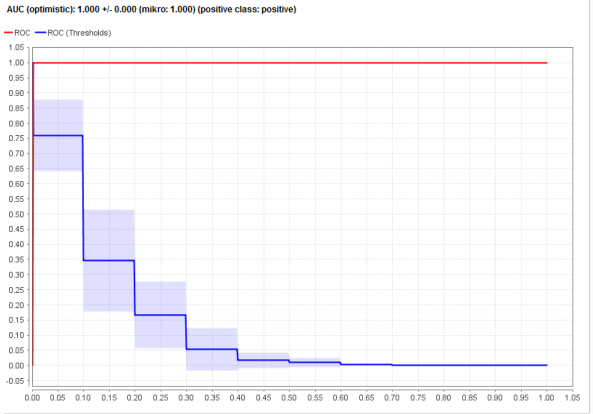
4.1 Uji coba Naive Bayes

Dengan menggunakan Rapid Miner, metrik yang tepat dari proses penambangan data tercapai. Hasil uji menggunakan model Naive Bayes di 90 % training dan 10 % uji dari total 200.

Tabel 2. Uji Model Confusion Matrix

accuracy: 100.00%			
	True negative	True positive	Class Precission
Pred. negative	100	0	100%
Pred. positive	0	100	100%
Class recall	100%	100%	

Menurut tabel diatas dari 200 titik data Twitter terkait film, 90% ditetapkan sebagai data training dan 10% data pengujian, atau 20 titik data. Hasil uji ditunjukkan pada tabel di atas. Karena ada 100 negatif dan 100 prediksi positif, hasilnya seperti yang diharapkan. Yaitu 100.00%. Berikut perhitungan yang divisualisasikan menggunakan kurva ROC dalam gambar kurva ROC naive bayes:



Gambar 2. Kurva ROC Naive Bayes

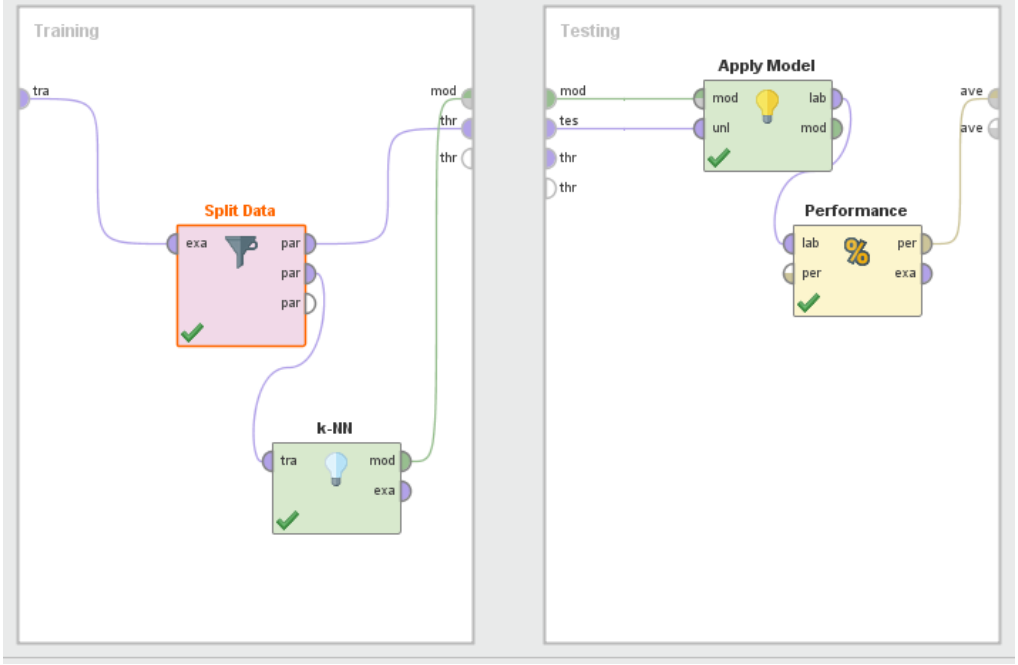
Adapun seluruh hasil pelatihan dan pengujian data dengan Naive Bayes, adalah sebagai berikut:

Tabel 3. Uji Naive Bayes

Uji Split Data	akurasi	AUC
Uji coba 1 (Training 90% dan Testing 10 %)	100%	1.000
Uji coba 2 (Training 80% dan Testing 20 %)	100%	1.000
Uji coba 3 (Training 70% dan Testing 30 %)	99.00%	1.000
Uji coba 4 (Training 60% dan Testing 40 %)	99.50%	1.000
Rata-Rata	99,63%	1.000

4.2 Uji Coba k-Nearest Neighbors

Berikut Analisis Sentimen Terhadap Opini Sinema Di Twitter, algoritma k-Nearest Neighbors digunakan bersamaan dengan database Rapid Miner. Data dibagi menjadi set pelatihan dan pengujian menggunakan model berikut:



Gambar 3. Model Pengujian Validasi K-nn

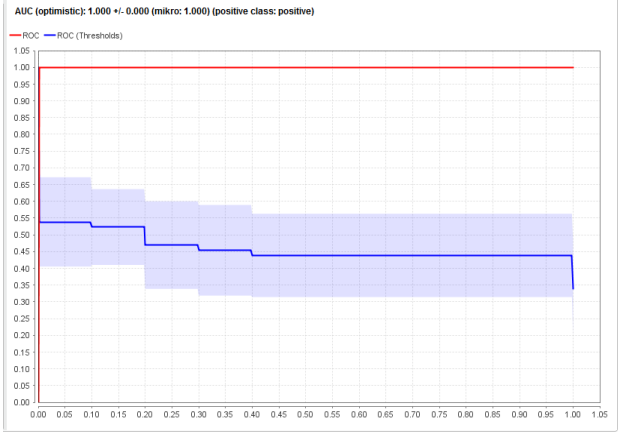
Karena metrik akurasi dari data pelatihan dapat dipecah menggunakan Rapid Miner, metrik akurasi sangat sensitif terhadap hasil analisis. Namun, hasil dari upaya tersebut menjadi semakin buruk karena nilai k yang digunakan meningkat. Hal ini disebabkan meningkatnya jumlah nilai terdekat yang digunakan dalam proses klasifikasi dan dapat terjaidngna noise. Rumus paling praktis untuk menentukan nilai k adalah k $\sqrt{}$ n, yang diterapkan pada kuantitas data pelatihan karena berfungsi sebagai contoh untuk data pengujian.

Memanfaatkan Rapid Miner, metrik akurasi dari proses pemisahan data tercapai. Hasil percobaan menggunakan model k-Nearest Neighbours untuk data training 90% dan data testing 10% dari total 200 data ditampilkan pada tabel berikut:

Tabel 4. Hasil Validasi k-NN

accuracy: 100.00%			
	True negative	True positive	Class Precission
Pred. negative	100	0	100.00%
Pred. positive	0	100	100.00%
Class recall	100%	100%	

Menurut Tabel 4, dari 200 pengguna Opini film di Twitter, 90% ditetapkan sebagai data pelatihan, dan 10% ditetapkan sebagai data pengujian, atau sekitar 20 pengguna. Sebutkan nilai K=n dari pelatihan data 180 hari sebagai 13. Hasil pengujian ditunjukkan pada tabel Confusion Matrix di atas. Jika 100 orang memprediksi hasil negatif, hasilnya seperti yang diharapkan negatif, dan jika 100 orang memprediksi hasil positif, hasilnya seperti yang diharapkan positif. Akurasi yang didapatkan pada k-NN adalah 100% dengan rumusan yaitu TP = 100 FP = 0 TN = 100 FN = 0. Akurasi sama dengan (TP + TN) / (TP + TN + FP + FN) = (100 + 100) / (100 + 100 + 0 + 0) = 200 / 200 = 1.



Gambar 4. Kurva ROC k-NN

Berikut hasil dari grafik ROC yang dihasilkan memiliki nilai AUC sebesar 1.000 dengan hasilnya *Excellent Classification*.

Tabel 5. Uji k-NN

Uji Split Data Pengujian	Akurasi	AUC
Uji coba 1 (Training 90% dan Testing 10 %)	100.00%	1.000
Uji coba 2 (Training 80% dan Testing 20 %)	99.00%	1.000
Uji coba 3 (Training 70% dan Testing 30 %)	99.50%	1.000
Uji coba 4 (Training 60% dan Testing 40 %)	98.50%	1.000
Rata-Rata	99,25%	1.000

4.3 Analisa Hasil Komparasi

Berdasarkan hasil percobaan yang dilakukan untuk menjawab permasalahan dengan prediksi hasil Opini Film di Twitter, dapat disimpulkan bahwa algoritma k-nn memberikan rata-rata akurasi nilai sebanyak 99,25% dan AUC sebesar 1.000 unutk nilai -rata -rata, sedangkan metode Naive Bayes memberikan hasil dengan nilai sebesar 99,63% dan nilai AUC sebesar 1.000. Dalam penelitian ini, dilakukan perbandingan dua metode klasifikasi yang lebih akurat yang digunakan untuk memprediksi opini film diTwitter.

Tabel 6. Hasil perbandingan keseluruhan

No	Percobaan	Naïve Bayes		k-NN	
		Accuracy	AUC	Accuracy	AUC
1	Percobaan 1 (Training 90% & Testing 10 %)	100%	1.000	100%	1.000
2	Percobaan 2 (Training 80% & Testing 20 %)	100%	1.000	99.00%	1.000
3	Percobaan 3 (Training 70% & Testing 30 %)	99.00%	1.000	99.50%	1.000
4	Percobaan 4 (Training 60% & Testing 40 %)	99.50%	1.000	98.50%	1.000
Rata-Rata		99,63%	1.000	99,25%	1.000

Tabel 7. Rata-rata perbandingan

Algoritma	Accuracy	AUC
Naive Bayes	99,63%	1.000
K-NN	99,25%	1.000

Dengan menggunakan data dari pengguna Twitter tentang Tweet tentang film dengan teks bahasa Indonesia, hasil Tabel 8 menunjukkan bahwa Algoritma Naive Bayes memiliki nilai lebih akurat dari Algoritma K-NN, dengan perbedaan akurasi sekitar 0,38% dan AUC nilai sekitar 1.000.

5. KESIMPULAN DAN SARAN

Dua algoritma data mining yang dibandingkan dalam analisis penelitian ini terhadap sentimen terhadap opini film berbahasa Indonesia di Twitter. Proses evaluasi dan divalidasi Pada algoritma tersebut menunjukkan hasil bahwa algoritma Naive Bayes memiliki lebih tinggi rasio keakuratan daripada algoritma k-NN perbedaan sebesar 0,38% dan tidak terlalu signifikan. Hasilnya, algoritme Naive Bayes dapat memprediksi opini Tweet film berbahasa Indonesia dengan lebih akurat dan presisi dibandingkan algoritme k-NN yang memiliki rata-rata hanya 99,25% dan AUC hanya 1.000. Hal ini didukung oleh akurasi algoritme yang sangat tinggi, yang ditunjukkan oleh AUC-nya sebesar 1.000 dan AUC sebesar 99,63%.

Ada beberapa hal yang dapat mempengaruhi situasi ini, seperti jumlah kumpulan data yang digunakan atau jumlah atribut yang digunakan, keduanya akan meningkatkan hasil. Oleh karena itu, hipotesis berdasarkan penelitian yang telah dilakukan tentang prediksi review film di laman Twitter berbahasa Indonesia menemukan bahwa algoritma dengan hasil lebih baik adalah Naive Bayes daripada algoritma K-nn saat memprediksi review film di Twitter. Sebagai aturan umum, saran dari penelitian ini adalah perlu mencoba

menggunakan algoritma klasifikasi lain untuk menentukan pendapat film, serta berbagai teknik pengoptimalan untuk meningkatkan skor, seperti adaboost.

DAFTAR PUSTAKA

[1] A. R. T. Lestari, R. S. Perdana, and M. A. Fauzi, "Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 12, pp. 1718–1724, 2017, [Online]. Available: http://j-ptiik.ub.ac.id.

[2] R. Apriani *et al.*, "Analisis Sentimen dengan Naïve Bayes Terhadap Komentar Aplikasi Tokopedia," *J. Rekayasa Teknol. Nusa Putra*, vol. 6, no. 1, pp. 54–62, 2019, [Online]. Available: https://rekayasa.nusaputra.ac.id/article/view/86.

[3] F. Rahutomo, P. Y. Saputra, and M. A. Fidyawan, "Implementasi Twitter Sentiment Analysis Untuk Review Film MenggunaRahutomo, F., Saputra, P. Y. and Fidyawan, M. A. (2018) 'Implementasi Twitter Sentiment Analysis Untuk Review Film Menggunakan Algoritma Support Vector Machine', Jurnal Informatika Polinema," *J. Inform. Polinema*, vol. 4, no. 2, p. 93, 2018.

[4] U. U. Amelya and R. K. Serli, "Analisa Minat Pelanggan Terhadap Produk Skincare Msglowskincare Nina Depok Dengan Algoritma Apriori," vol. 8, no. 2, pp. 187–193, 2022.

[5] D. J. Lubis and G. K. Gusti, "Penerapan Algoritma Naïve Bayes Untuk Penentuan Balita Penerima Makanan Tambahan (PMT) Berdasarkan Status Gizi Di Pos Pelayanan Terpadu (POSYANDU)," vol. 13, no. 1, pp. 58–66, 2023, doi: 10.36350/jbs.v13i1.177.

[6] A. Purwanto and H. W. Nugroho, "Analisa Perbandingan Kinerja Algoritma C4.5 Dan Algoritma K-Nearest Neighbors Untuk Klasifikasi Penerima Beasiswa," *J. Teknoinfo*, vol. 17, no. 1, p. 236, 2023, doi: 10.33365/jti.v17i1.2370.