Jurnal Dinamika Informatika
Volume 15, No 1, Maret 2026
ISSN 1978-1660 : 112-122
ISSN *online* 2549-8517

Perbandingan Kinerja Naive Bayes dan KNN dalam Klasifikasi Sentimen Ulasan Film Horor

Cantriya Anastasya Simbolon[1], Maria Angelina Lubis[2], Sardo Pardingotan Sipayung[3]

[1,2,3]Jurusan Teknik Informatika, Universitas Katolik Santo Thomas

e-mail: [1]cantriyasimbolonbelajar@gmai;.com , [2] mariaangelinalubis@gmail.com, 3pinsarsiphom@gmail.com

Intisari

Lonjakan ulasan film horor di platform digital memerlukan sistem klasifikasi otomatis untuk memahami sentimen penonton secara efisien. Penelitian ini bertujuan membandingkan kinerja algoritma Naive Bayes dan *K-Nearest Neighbors* (KNN) dalam mengklasifikasikan sentimen ulasan film horor berbahasa Inggris. Metodologi penelitian melibatkan pengolahan 3.000 data dari Kaggle menggunakan perangkat lunak RapidMiner, dengan tahapan pra-pemrosesan meliputi pembobotan TF-IDF, *tokenization*, *filtering*, dan *stemming*. Pengujian dilakukan melalui skema 10-*fold cross validation* untuk menjamin stabilitas hasil. Temuan penelitian menunjukkan perbedaan performa yang signifikan, di mana Naive Bayes meraih akurasi sebesar 88,53%, jauh mengungguli KNN yang hanya mencapai 40,47%. Rendahnya akurasi KNN disebabkan oleh kompleksitas perhitungan jarak pada data teks berdimensi tinggi. Disimpulkan bahwa Naive Bayes merupakan model yang lebih reliabel dan efektif untuk klasifikasi sentimen ulasan film horor. Hasil ini memberikan kontribusi berupa rekomendasi algoritma optimal bagi pengembangan sistem analisis opini otomatis.

Kata kunci: Analisis Sentimen, Naive Bayes, K-Nearest Neighbors, RapidMiner, Film Horor

Abstract

The surge in horror movie reviews on digital platforms requires an automated classification system to efficiently understand audience sentiment. This study aims to compare the performance of Naive Bayes and K-Nearest Neighbors (KNN) algorithms in classifying the sentiment of English-language horror movie reviews. The research methodology involved processing 3,000 data points from Kaggle using RapidMiner software, with pre-processing stages including TF-IDF weighting, tokenization, filtering, and stemming. Testing was conducted using a 10-fold cross-validation scheme to ensure the stability of the results. The research findings show a significant difference in performance, with Naive Bayes achieving an accuracy of 88.53%, far surpassing KNN, which only reached 40.47%. The low accuracy of KNN is due to the complexity of distance calculations in high-dimensional text data. It is concluded that Naive Bayes is a more reliable and effective model for classifying the sentiment of horror movie reviews. These results contribute to recommendations for optimal algorithms for the development of automatic opinion analysis systems.

Keywords: Sentiment Analysis, Naive Bayes, K-Nearest Neighbors, RapidMiner,Horror Movies.

PENDAHULUAN

Transformasi digital dalam industri hiburan telah memicu pergeseran cara audiens memberikan umpan balik, di mana ulasan daring kini menjadi determinan utama dalam keberhasilan sebuah karya sinematografi. IMDb (Internet Movie Database) merupakan salah satu sumber data teks yang sangat populer[1].Platform pemeringkatan film global seperti IMDb menghasilkan ribuan ulasan setiap harinya yang mencerminkan persepsi emosional penonton secara eksplisit. Khusus pada genre film horor, opini yang diberikan sering kali mengandung diksi yang subjektif dan intens, sehingga menjadi sumber data yang sangat kaya untuk dianalisis. Namun, ketersediaan data masif yang tersedia di berbagai repositori terbuka seperti Kaggle mustahil diproses secara manual tanpa bantuan otomasi. **Kaggle** adalah platform terkemuka yang memungkinkan akses ke berbagai jenis dataset untuk penelitian, pembelajaran mesin, dan proyek data science [2]. Oleh karena itu, klasifikasi sentimen menggunakan teknik *data mining* menjadi instrumen krusial untuk mengekstraksi informasi berharga dari teks yang tidak terstruktur

Tantangan utama dalam analisis sentimen ulasan film berbahasa Inggris terletak pada kompleksitas leksikon, penggunaan ironi, serta struktur kalimat yang tidak konsisten. Persoalan ini menuntut tahapan pra-pemrosesan teks yang presisi agar model klasifikasi mampu mengenali pola opini dengan akurat. Penggunaan perangkat lunak pengolah data berbasis visual seperti **RapidMiner** memberikan keunggulan dalam merancang alur kerja pemrosesan teks yang sistematis, mulai dari pembersihan data hingga pembobotan fitur menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF). Pemilihan algoritma klasifikasi yang diintegrasikan ke dalam lingkungan tersebut menjadi faktor penentu reliabilitas hasil penelitian.

Dalam literatur pembelajaran mesin, dua algoritma yang sering diperbandingkan kinerjanya adalah Naive Bayes dan K-Nearest Neighbors (KNN). Naive Bayes sangat dikenal karena efisiensinya dalam menangani data berdimensi tinggi melalui pendekatan probabilistik [3]). Di sisi lain, KNN menawarkan pendekatan berbasis jarak yang lebih fleksibel terhadap distribusi data. Perbedaan fundamental dalam mekanisme kerja kedua algoritma ini memunculkan urgensi untuk melakukan pengujian komparatif guna menentukan model mana yang lebih tangguh dalam menangani data ulasan film yang bersifat jarang (*sparse*). Naive bayes classifer merupakan Metode klasifikasi yang memiliki beberapa fase penyelesaian yaitu dimulai dari Training Data, Learning Algorithm, Model, Test Data dan diakhiri dengan proses Testing sehingga dihasilkan sebuah keputusan yang akurat[4]. KNN mengklasifikasikan objek berdasarkan data pembelajaran yang memiliki jarak terdekat dengan objek tertentu. Algoritma Nearest Neighbor hanya mempertimbangkan satu data pembelajaran terdekat sebagai basis klasifikasi[5].

Metode TF-IDF menggabungkan dua konsep, yaitu frekuensi kata dalam dokumen dan pengacakan dokumen yang berisi kata[6].Kata yang sering muncul di satu dokumen namun jarang muncul di dokumen lain akan diberikan bobot yang lebih tinggi karena dianggap lebih representatif sebagai ciri khas dokumen tersebut.

Jurnal Dinamika Informatika
Volume 15, No 1, Maret 2026
ISSN 1978-1660 : 112-122
ISSN *online* 2549-8517

Naive Bayes classifier (NBC) ialah salah satu metode klasifikasi yang sering digunakan[7].Dalam klasifikasi teks, algoritma ini menghitung probabilitas kemunculan sekumpulan kata terhadap label tertentu. Keunggulannya terletak pada kecepatan komputasi dan efektivitasnya dalam menangani data dengan jumlah fitur yang sangat banyak.

K-Nearest Neighbor merupakan teknik mencari anggota k target dalam data latih (training) yang terdekat dengan target pada data testing atau data baru[8].Penentuan "kedekatan" ini umumnya dihitung menggunakan metrik jarak, seperti *Euclidean Distance*. KNN sangat sensitif terhadap dimensi data yang besar karena semakin banyak fitur, maka perhitungan jarak menjadi semakin kompleks.

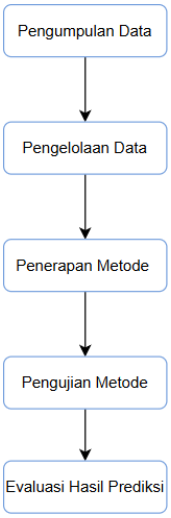
Sejumlah studi literatur sebelumnya telah mencoba membedah efektivitas kedua metode ini dalam berbagai domain kasus. Penelitian yang dilakukan oleh [9]mengenai klasifikasi judul artikel pada jurnal ilmiah menunjukkan adanya kompetisi performa yang ketat antara kedua algoritma tersebut. Dalam studi lain yang dilakukan oleh [10]ditemukan bahwa metode K-Nearest Neighbor yang diintegrasikan dengan fitur HSV dan teknik *cropping* mampu menghasilkan akurasi yang sangat tinggi untuk klasifikasi gender. Hal ini mengindikasikan bahwa KNN memiliki keunggulan pada data yang memiliki karakteristik visual atau spasial yang kuat.Namun, temuan pada domain data numerik menunjukkan hasil yang berbeda. [11]dalam
penelitiannya terkait klasifikasi pemberian pinjaman nasabah melaporkan bahwa algoritma KNN dan Naive Bayes memiliki nilai akurasi yang identik, yakni sebesar 77%, yang keduanya mengungguli performa *Support Vector Machine* (SVM) yang hanya mencapai 50%. Sementara itu, pada ranah medis, [12] melalui analisis diagnosis penyakit diabetes mellitus menegaskan bahwa model algoritma Naive Bayes memiliki tingkat akurasi yang lebih superior dibandingkan dengan model KNN.Keragaman hasil dari berbagai penelitian di atas menunjukkan bahwa performa algoritma sangat bergantung pada karakteristik *dataset* yang digunakan.

Meskipun klasifikasi sentimen telah banyak dikaji, penelitian yang secara spesifik membandingkan kinerja Naive Bayes dan KNN pada domain ulasan film horor dengan prosedur validasi yang ketat masih relatif terbatas. Celah penelitian inilah yang memotivasi dilakukannya studi ini dengan memanfaatkan dataset ulasan IMDb untuk menyediakan bukti empiris mengenai algoritma mana yang paling optimal.

Berdasarkan latar belakang di atas, penelitian ini bertujuan untuk mengevaluasi dan membandingkan kinerja algoritma Naive Bayes dan K-Nearest Neighbors (KNN) menggunakan perangkat lunak **RapidMiner**. RapidMiner merupakan perangkat lunak yang bersifat terbuka (open source) yang diciptakan dengan menggunakan bahasa pemrograman java sehingga bisa diakses oleh semua sistem operasi[13]. Penelitian ini secara sistematis melakukan tahapan pra-pemrosesan teks, melakukan pengujian model dengan skema 10-fold *cross validation*, serta menganalisis performa berdasarkan metrik *accuracy, precision, recall*, dan *F1-score*. Hasil penelitian diharapkan dapat menjadi referensi akademik yang valid di bidang *natural language processing*.

METODE PENELITIAN

Metodologi penelitian merupakan suatu cara atau strategi sistematis yang digunakan oleh peneliti dalam melakukan kegiatan penelitian[14]. Tahapan penelitian ini disusun secara terstruktur untuk menjamin validitas hasil dalam membandingkan performa algoritma *Naive Bayes* dan *K-Nearest Neighbors* (KNN). Alur kerja penelitian ini secara garis besar mengikuti prosedur yang ditunjukkan pada Gambar 1, mulai dari tahap pengumpulan data hingga evaluasi hasil prediksi.



Gambar 1. Alur Proses Metode Penelitian

1. Pengumpulan Data

   Pengumpulan data merupakan langkah awal yang penting dalam penelitian serta pengambilan keputusan[15]. Dataset terdiri dari 3.000 data ulasan yang mencakup atribut kunci seperti review_id, movie_title, review_text, rating, review_date, dan sentiment_label. Data ini diimpor ke dalam lingkungan kerja menggunakan operator Read CSV untuk pemrosesan lebih lanjut.
2. Pengelolaan Data (*Data Preprocessing*)

   Tahap pengelolaan data merupakan langkah krusial untuk mentransformasi teks tidak terstruktur menjadi bentuk numerik agar dapat diproses oleh algoritma pembelajaran mesin. Proses ini dilakukan melalui dua tahapan utama pada Altair AI Studio:
   a. Persiapan Atribut: Sebelum pengolahan teks, dilakukan perubahan tipe data menggunakan operator Nominal to Text pada atribut ulasan. Selanjutnya, operator Set Role digunakan untuk menetapkan atribut sentiment_label sebagai target klasifikasi (*label*).
   b. Pemrosesan Teks (*Text Processing*): Menggunakan operator Process Documents from Data dengan metode pembobotan TF-IDF (*Term Frequency-Inverse Document Frequency*) untuk pembentukan vektor kata. Di dalam sub-proses ini, dilakukan serangkaian langkah pembersihan teks yang meliputi:

Jurnal Dinamika Informatika
Volume 15, No 1, Maret 2026
ISSN 1978-1660 : 112-122
ISSN *online* 2549-8517

1) Tokenization: Memecah kalimat ulasan menjadi satuan kata tunggal (*token*).
2) Filter Stopwords (English): Menghapus kata-kata umum dalam bahasa Inggris yang tidak memiliki bobot informatif bagi klasifikasi sentimen.
3) Stem (Porter): Mengubah setiap kata ke bentuk dasarnya menggunakan algoritma Porter untuk mengurangi variasi morfologi kata.

3. Penerapan dan Pengujian Metode

Penelitian ini membandingkan dua algoritma klasifikasi, yaitu Naive Bayes dan K-Nearest Neighbors (KNN). Untuk menghasilkan estimasi performa yang stabil dan menghindari bias pada pembagian data, pengujian dilakukan menggunakan operator Cross Validation. Naïve Bayes Classifier mengadopsi ilmu statistika yaitu dengan menggunakan teori kemungkinan (Probabilitas) untuk menyelesaikan sebuah kasus Supervised Learning, artinya dalam himpunan data terdapat Label, Class atau Target sebagai acuan[16]. K-Nearest Neigbors (KNN) adalah sebuah algoritma klasifikasi popular yang sering digunakan dalam machine learning[17].

Konfigurasi pengujian menggunakan skema 10-Fold Cross Validation dengan metode pengambilan sampel Stratified Sampling. Dalam skema ini, dataset dibagi menjadi sepuluh bagian seimbang, di mana setiap iterasi menggunakan sembilan bagian sebagai data latih (*training*) di sisi kiri proses, dan satu bagian sebagai data uji (*testing*) di sisi kanan proses menggunakan operator Apply Model.

4. Evaluasi Hasil Prediksi

Tahap akhir dari metodologi ini adalah mengukur kinerja dari kedua model klasifikasi menggunakan operator Performance. Evaluasi dilakukan dengan menganalisis hasil *Confusion Matrix* untuk mendapatkan nilai parameter keberhasilan penelitian, yang meliputi nilai Accuracy, Precision, Recall, dan F1-Score. Hasil perbandingan metrik-metrik tersebut akan menentukan algoritma mana yang memiliki efektivitas tertinggi dalam mengklasifikasikan sentimen ulasan film horor.

HASIL DAN PEMBAHASAN

Hasil Bagian ini menguraikan hasil implementasi eksperimen klasifikasi sentimen menggunakan dua algoritma berbeda. Seluruh tahapan pengujian dilakukan secara sistematis mulai dari penyiapan data hingga perolehan metrik akurasi akhir.

1. Analisis Desain Proses dan Dataset

Pada gambar 2, ditampilkan hasil import klasifikasi *dataset* ke dalam lingkungan kerja RapidMiner. Dataset yang telah diimpor dipastikan memiliki atribut yang sesuai dengan peran masing-masing, terutama penetapan label sentimen yang akan diprediksi.

	review_id *polynominal*	movie_title *polynominal*	review_text *polynominal*	rating *integer*	review_date *date*	sentiment_label *polynominal*
1	R001678	The Conjuring	At first I thought somewha...	10	Jan 3, 2015	positive
2	R001224	The Conjuring	To be fair, extremely bore...	7	Jan 3, 2015	positive
3	R002965	Get Out	At first I thought terrified a...	9	Jan 4, 2015	positive
4	R000568	Midsommar	Right from the start, a bit ...	4	Jan 5, 2015	negative
5	R001136	It Follows	I've never felt very intrigue...	8	Jan 7, 2015	positive
6	R002753	Paranormal Activity	Honestly, a bit disturbed a...	5	Jan 9, 2015	neutral
7	R001206	The Babadook	Right from the start, unex...	6	Jan 10, 2015	neutral
8	R002010	Paranormal Activity	This movie left me under...	8	Jan 14, 2015	positive
9	R000701	Paranormal Activity	Right from the start, som...	5	Jan 15, 2015	neutral
10	R002124	The Witch	Honestly, unexpectedly b...	6	Jan 21, 2015	neutral
11	R001179	The Ring	I've never felt a bit shocke...	4	Jan 22, 2015	negative
12	R002949	Paranormal Activity	I went in expecting some...	6	Jan 23, 2015	neutral
13	R001532	The Babadook	At first I thought extremely...	2	Jan 23, 2015	negative
14	R000702	The Witch	Right from the start, extre...	9	Jan 24, 2015	positive
15	R000521	The Exorcist	I went in expecting some...	3	Jan 24, 2015	negative
16	R002942	Paranormal Activity	To be fair, very shocked a...	5	Jan 25, 2015	neutral
17	R002144	Halloween	This movie left me extrem...	7	Jan 27, 2015	positive
18	R000322	The Babadook	I've never felt impressed a...	8	Jan 28, 2015	positive

Gambar 2. Dataset yang akan diolah

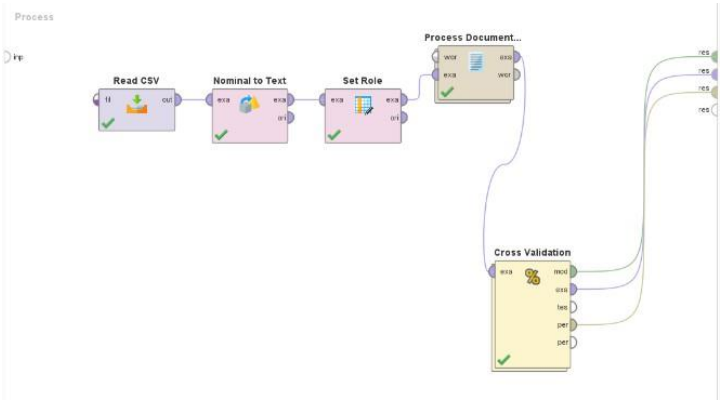
2. Implementasi Pemodelan Naive Bayes dan KNN

Eksperimen klasifikasi dilakukan dengan merancang alur kerja paralel untuk membandingkan efektivitas kedua algoritma secara objektif. Gambar 3 menunjukkan desain proses utama di RapidMiner yang digunakan secara seragam baik untuk pengujian Naive Bayes maupun KNN. Tahapan ini dimulai dari pembacaan data, pengaturan *role*, hingga penggunaan operator *Process Documents from Data*. Di dalam operator tersebut, diterapkan pembobotan TF-IDF serta teknik *Tokenization*, *Filtering Stopwords* (English), dan *Stemming* (Porter) guna mengoptimalkan representasi kata dari ulasan film horor sebelum masuk ke tahap pemodelan.

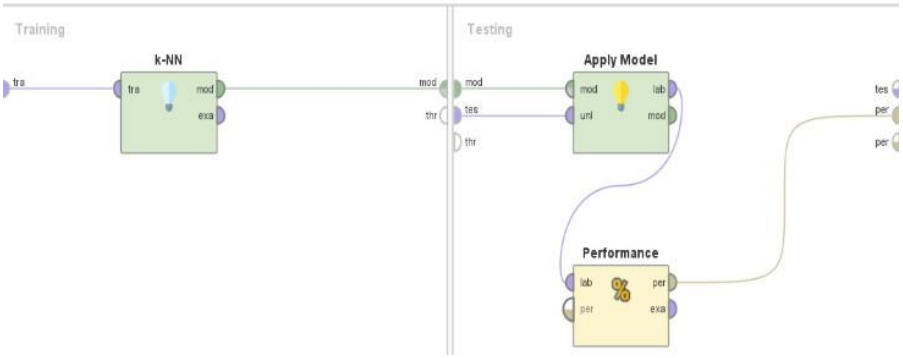
Perbedaan utama dalam eksperimen ini terletak pada fase pembelajaran di dalam operator *Cross Validation*. Gambar 4 memperlihatkan konfigurasi pada sisi *training* di mana algoritma K-Nearest Neighbors (KNN) diimplementasikan. Skema pengujian menggunakan 10-Fold Cross Validation dengan metode *stratified sampling* untuk memastikan distribusi label yang seimbang di setiap lipatan (*fold*).

Sementara itu, untuk pengujian kedua, sisi *training* dikonfigurasi menggunakan algoritma Naive Bayes sebagaimana ditunjukkan pada Gambar 5. Kedua model kemudian diuji pada sisi *testing* menggunakan operator *Apply Model* dan dievaluasi kinerjanya melalui operator *Performance* untuk menghasilkan metrik akurasi yang akurat dan stabil.

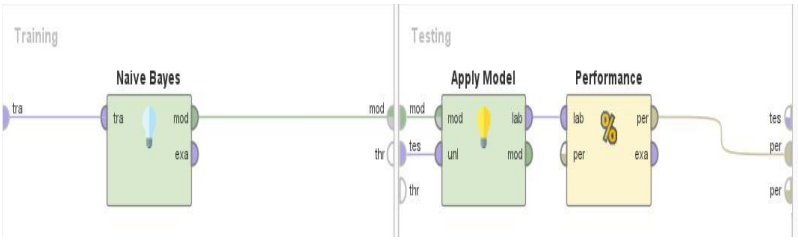
Jurnal Dinamika Informatika
Volume 15, No 1, Maret 2026
ISSN 1978-1660 : 112-122
ISSN *online* 2549-8517



Gambar 3. Desain Proses Utama Pengolahan Data Teks dan Alur *Cross Validation* di RapidMiner.



Gambar 4. Implementasi Algoritma K-Nearest Neighbors (KNN) pada Sisi *Training* dalam Proses *Cross Validation*.



Gambar 5. Implementasi Algoritma Naive Bayes pada Sisi *Training* dalam Proses *Cross Validation*.

Eksperimen dilakukan dengan menjalankan dua skema pengujian paralel untuk membandingkan efektivitas algoritma. Gambar 4 menunjukkan detail desain pengujian data menggunakan RapidMiner untuk algoritma Naive Bayes. Proses ini melibatkan operator *Process Documents from Data* yang mengimplementasikan pembobotan TF-IDF serta teknik *Tokenization*, *Filtering Stopwords*, dan *Stemming* (Porter) untuk mengoptimalkan representasi kata.

Di sisi lain, pengujian data menggunakan algoritma K-Nearest Neighbors (KNN) ditampilkan pada Gambar 5. Serupa dengan model pertama, pengujian ini juga menggunakan skema 10-Fold Cross Validation dengan tipe

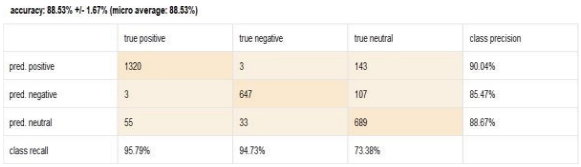
pengambilan sampel *stratified sampling* untuk menjamin bahwa hasil yang diperoleh bersifat objektif dan stabil.

3. Perbandingan Akurasi dan Evaluasi Model

accuracy: 40.47% +/- 3.08% (micro average: 40.47%)

	true positive	true negative	true neutral	class precision
pred. positive	801	340	538	47.71%
pred. negative	221	156	144	29.94%
pred. neutral	356	187	257	32.12%
class recall	58.13%	22.84%	27.37%	

Gambar 6. Hasil *Confusion Matrix* dan Nilai Akurasi Algoritma *K-Nearest Neighbors* (KNN).



accuracy: 88.53% +/- 1.67% (micro average: 88.53%)

	true positive	true negative	true neutral	class precision
pred. positive	1320	3	143	90.04%
pred. negative	3	647	107	85.47%
pred. neutral	55	33	689	88.67%
class recall	95.79%	94.73%	73.38%	

Gambar 7. Hasil *Confusion Matrix* dan Nilai Akurasi Algoritma *Naive Bayes*.

Setelah seluruh proses pengujian selesai dijalankan, hasil evaluasi menunjukkan perbedaan performa yang sangat kontras antara kedua algoritma dalam mengklasifikasikan sentimen ulasan film horor. Berikut adalah rincian capaian kinerjanya:

a. **Algoritma K-Nearest Neighbors (KNN)** Berdasarkan hasil pengujian pada **Gambar 6**, model KNN menghasilkan tingkat akurasi yang rendah, yaitu sebesar **40,47% (+/- 3,08%)**. Rendahnya akurasi ini disebabkan oleh banyaknya kesalahan klasifikasi pada seluruh kategori sentimen. Hal ini terlihat jelas pada nilai *recall* untuk kelas negatif yang hanya mencapai **22,84%**, menunjukkan bahwa model mengalami kesulitan dalam mengidentifikasi ulasan yang bersifat kritik atau tidak puas.
b. **Algoritma Naive Bayes** Sebaliknya, sebagaimana ditunjukkan pada **Gambar 7**, algoritma Naive Bayes menunjukkan performa yang jauh lebih unggul dengan tingkat akurasi mencapai **88,53% (+/- 1,67%)**. Model ini memiliki kemampuan yang sangat baik dalam mengenali pola teks ulasan, terutama pada kelas positif dengan nilai *recall* tertinggi sebesar **95,79%**. Stabilitas akurasi yang tinggi ini membuktikan efektivitas pendekatan probabilistik dalam menangani data ulasan film yang kompleks.

4. Pembahasan

Melalui perbandingan kedua hasil tersebut, algoritma **Naive Bayes** terbukti jauh lebih unggul dibandingkan **KNN** dalam domain analisis sentimen ini. Tingginya akurasi Naive Bayes (88,53%) mengonfirmasi hipotesis bahwa model probabilistik lebih efektif dalam menangani data teks yang bersifat *sparse* dan berdimensi tinggi setelah proses TF-IDF. Sebaliknya, KNN dengan akurasi 40,47% mengalami kesulitan karena ketergantungannya pada perhitungan jarak antar fitur teks yang sangat luas, sehingga sering terjadi tumpang tindih (*overlap*) antar kelas sentimen yang berbeda.

Jurnal Dinamika Informatika
Volume 15, No 1, Maret 2026
ISSN 1978-1660 : 112-122
ISSN *online* 2549-8517

KESIMPULAN

Berdasarkan rangkaian eksperimen yang telah dilaksanakan dalam penelitian ini, dapat ditarik kesimpulan bahwa pemilihan algoritma klasifikasi memiliki dampak yang sangat signifikan terhadap akurasi analisis sentimen pada ulasan film horor berbahasa Inggris. Melalui pengujian menggunakan skema *10-fold cross validation* pada dataset yang berjumlah 3.000 data, hasil penelitian menunjukkan perbedaan performa yang sangat kontras antara kedua metode yang diuji.

Algoritma **Naive Bayes** terbukti menjadi model yang paling optimal dengan perolehan nilai akurasi mencapai **88,53%**. Keunggulan Naive Bayes dalam menangani fitur teks yang bersifat *sparse* (jarang) setelah proses pembobotan TF-IDF menunjukkan bahwa pendekatan probabilistik sangat relevan untuk tugas klasifikasi teks berdimensi tinggi. Di sisi lain, algoritma **K-Nearest Neighbors (KNN)** memberikan hasil yang kurang memuaskan dengan tingkat akurasi sebesar **40,47%**. Rendahnya performa KNN disebabkan oleh kompleksitas perhitungan jarak pada fitur teks yang sangat luas, sehingga model mengalami kesulitan dalam membedakan batas-batas kelas sentimen secara presisi.

Penelitian ini memberikan kontribusi berupa rekomendasi metodologis bahwa untuk dataset ulasan film horor dari IMDb, Naive Bayes merupakan algoritma yang jauh lebih reliabel dibandingkan KNN. Untuk pengembangan penelitian selanjutnya, disarankan untuk melakukan optimasi pada parameter algoritma KNN atau mencoba integrasi metode *ensemble learning* guna mengeksplorasi potensi peningkatan akurasi pada domain data yang serupa.

SARAN

Berdasarkan keterbatasan dan hasil yang diperoleh dalam penelitian ini, peneliti merumuskan beberapa saran untuk pengembangan studi di masa mendatang:

1. **Optimasi Hyperparameter dan Metrik Jarak:** Mengingat rendahnya performa algoritma *K-Nearest Neighbors* (KNN) pada data teks penelitian ini, disarankan bagi peneliti selanjutnya untuk melakukan eksperimen lebih mendalam terhadap optimasi nilai k melalui metode *grid search*. Selain itu, penggunaan metrik jarak alternatif seperti *Cosine Similarity* perlu dipertimbangkan, mengingat metrik tersebut cenderung lebih tangguh dalam menangani data teks berdimensi tinggi dibandingkan *Euclidean Distance*.

2. **Penerapan Teknik Reduksi Dimensi:** Guna mengatasi fenomena *Curse of Dimensionality* yang menghambat kinerja KNN, penelitian berikutnya dapat mengintegrasikan teknik seleksi fitur seperti *Information Gain* atau *Chi-Square*. Selain itu, penggunaan metode ekstraksi fitur seperti *Principal Component Analysis* (PCA) atau *Latent Semantic Analysis* (LSA) dapat diterapkan untuk

menyederhanakan ruang fitur tanpa kehilangan informasi semantik yang signifikan.

3. **Eksplorasi Representasi Fitur Berbasis Semantik:** Disarankan untuk membandingkan pembobotan TF-IDF dengan teknik *word embedding* seperti Word2Vec atau FastText. Pendekatan ini diharapkan mampu menangkap konteks emosional dan hubungan antar kata pada ulasan film horor secara lebih mendalam dibandingkan pendekatan berbasis frekuensi kata tunggal.
4. **Implementasi Metode Ensemble Learning:** Untuk meningkatkan stabilitas akurasi pada domain data yang bersifat *sparse*, penelitian selanjutnya dapat mencoba integrasi metode *ensemble* seperti *Random Forest* atau *XGBoost* sebagai pembanding tambahan guna mengeksplorasi potensi peningkatan performa klasifikasi pada ulasan film.

DAFTAR PUSTAKA

[1] Yolanda Aprilia and Wiwin Widhihastuty, "ANALISIS SENTIMEN ULASAN FILM PADA IMDB MENGGUNAKAN ALGORITMA NAÏVE BAYES," *IKRAITH-INFORMATIKA* , vol. 10, pp. 1–7, 2024.

[2] Abdul Jalil, Ahmad Homaidi, and Zaehol Fatah, "Implementasi Algoritma Support Vector Machine Untuk Klasifikasi Status Stunting Pada Balita," *G-Tech : Jurnal Teknologi Terapan*, vol. 8, pp. 2070–2079, 2024.

[3] Pengming Hu, Weidong Yang, Xuyu Wang, and Shiwen Mao, "Contact-free wheat mildew detection with commodity wifi".

[4] Eko Martantoh and Nur Yanih, "Implementasi Metode Naïve Bayes Untuk Klasifikasi Karakteristik Kepribadiaan Siswa Di Sekolah MTS Darussa'adah Menggunakan PHP MySQL ," *JTSI*, vol. 3, pp. 166–175, 2022.

[5] Raja Sakti Arief Daulay, "Analisis Kritis dan Pengembangan Algoritma K-Nearest Neighbor (KNN): Sebuah Tinjauan Literatur," *Jurnal Pendidikan Sains dan Komputer*, vol. 4, pp. 131–141, 2024.

[6] Shalvan Chamira, "Implementasi Metode Text Mining Frequency-Invers Document Frequency (Tf-Idf) Untuk Monitoring Diskusi Online," *Journal of Informatics, Electrical and Electronics Engineering*, vol. 1, pp. 97–102, 2022.

[7] Angga Pebdika, Ruli Herdiana, and Dodi Solihudin, "KLASIFIKASI MENGGUNAKAN METODE NAIVE BAYES UNTUK MENENTUKAN CALON PENERIMA PIP," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, 2023.

[8] Anton Prasetiyo and Arita Witanti, "Implementasi Metode K-Nearest Neighbor Dalam Menentukan Waktu Optimal Penarikan Pesanan Driver Ojol," *Jurnal Teknik Informatika dan Sistem Informasi* , vol. 11, pp. 255–269, 2024.

[9] Anang Prayogo, Fauziah, and Winarsih, "PERBANDINGAN ALGORITMA NAÏVE BAYES DAN K-NEAREST NEIGHBOR PADA KLASIFIKASI JUDUL ARTIKEL PADA JURNAL ILMIAH," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 8, pp. 1327–1338, 2023.

[10] Calvin Kurniawan and Hafiz Irsyad, "Perbandingan Metode K-Nearest Neighbor Dan Naïve Bayes Untuk Klasifikasi Gender Berdasarkan Mata," *Jurnal Algoritme*, vol. 2, pp. 82–91, 2022.
[11] Z Saputra, H A Supahri, R Ismanizan, and Rahmaddeni, "Perbandingan Algoritma KNN (K-Nearest Neighbors), Naïve Bayes, Dan SVM (Support Vector Machine) Untuk Klasifikasi Pemberian Pinjaman Nasabah," *Jurnal Ilmiah Sistem Informasi dan Teknik Informatika (JISTI)* , vol. 7, pp. 67–75, 2024.
[12] Aryanto Bangun Widodo and Ericks Rachmat Swedia, "Analisis Perbandingan Algoritma KNN dan Naïve Bayes dalam Mendiagnosis Penyakit Diabetes Mellitus," *Jurnal Ilmiah KOMPUTAS*, vol. 2024, pp. 387–396, 23AD.
[13] Apriyadi, Muhammad Ridwan Lubis, and Bahrudi Efendi Damanik, "PENERAPAN ALGORITMA C5.0 DALAM MENENTUKAN TINGKAT PEMAHAMAN MAHASISWA TERHADAP PEMBELAJARAN DARING," *KOMPUTA : Jurnal Ilmiah Komputer dan Informatika*, vol. 11, pp. 11–20, 2022.
[14] Muammar Khaddafi, Laina Fitri, Putri Sarah, and Anis Shafa, "PENTINGNYA PEMILIHAN METODE PENELITIAN YANG TEPAT DALAM PENELITIAN ILMIAH," *Jurnal Intelek Insan Cendikia*, vol. 2, pp. 13372–13376, 2025.
[15] Bakhrudin All Habsy *et al.*, " Manajemen Pengumpulan Data," *Jurnal Mahasiswa Kreatif* , vol. 2, pp. 34–46, 2024.
[16] Eko Martantoh and Nur Yanih, "Implementasi Metode Naïve Bayes Untuk Klasifikasi Karakteristik Kepribadiaan Siswa Di Sekolah MTS Darussa'adah Menggunakan PHP MySQL," *JTSI*, vol. 3, pp. 166–175, 2022.
[17] Raja Sakti Arief Daulay, "Analisis Kritis dan Pengembangan Algoritma K-Nearest Neighbor (KNN): Sebuah Tinjauan Literatur," *Jurnal Pendidikan Sains dan Komputer*, vol. 2, pp. 131–141, 2024.