

# Analisis Sentimen Movie Review Berbasis Naïve Bayes Pada Dataset IMDb

Muh. Tigra Gaizka  
Fakultas Informatika  
Universitas Telkom  
Bandung, Indonesia

tigragaizka@student.telkomuniversity.ac.id

Dr. Kemas Muslim L.  
Fakultas Informatika  
Universitas Telkom  
Bandung, Indonesia

kemasmuslim@telkomuniversity.ac.id

Widi Astuti, S.T., M.Kom.  
Fakultas Informatika  
Universitas Telkom  
Bandung, Indonesia

widiwdu@telkomuniversity.ac.id

**Abstrak** — Analisis sentimen sangat penting untuk memahami opini dan reaksi pengguna terhadap produk atau layanan. Movie review memiliki kompleksitas yang mencakup berbagai aspek seperti plot, akting, dan visual. Memahami sentimen berdasarkan review ini dapat memberikan wawasan berharga bagi pembuat film, sutradara, dan pembuat kebijakan. Namun, analisis sentimen movie review tidak dapat dikatakan mudah karena tantangan-tantangan yang dapat dihadapi. Tantangan ini bisa berupa penggunaan bahasa review yang tidak terstruktur, pemilihan fitur-fitur yang tidak tepat, dataset yang kompleks, atau metode klasifikasi yang berat. Dalam penelitian ini, digunakan dataset IMDb yang berisi 50.000 data review serta label sentimen (positif, negatif). Sedangkan algoritma klasifikasi yang digunakan ialah Naïve Bayes, yang tepatnya ialah Multinomial Naïve Bayes. Algoritma ini didukung oleh vectorization menggunakan TF-IDF, dan feature selection menggunakan chi-squared. Hasil penelitian menunjukkan bahwa model Multinomial Naïve Bayes dengan parameter TF-IDF yang optimal dapat menghasilkan accuracy sebesar 86.14%. Performa model tersebut meningkat menjadi 87.71% setelah mengimplementasikan chi-squared dengan parameter yang juga optimal. Penelitian ini memberikan wawasan mengenai teknik serta tantangan dalam analisis sentimen movie review pada IMDb dengan metode klasifikasi Multinomial Naïve Bayes, serta menguji parameter pre-processing untuk menentukan dan mengevaluasi performa yang dihasilkan.

**Kata kunci**— Multinomial Naïve Bayes, TF-IDF, chi-squared, accuracy, IMDb

## I. PENDAHULUAN

Isi Analisis sentimen merupakan bidang studi yang menganalisis opini, sentimen, evaluasi, penilaian, tingkah laku, dan emosi masyarakat terhadap suatu produk, layanan, organisasi, individu, masalah, peristiwa, atau topik dari suatu hal [1]. Pada masa digital seperti sekarang ini, analisis sentimen menjadi sangat relevan dalam berbagai macam bidang, seperti penilaian produk, peringkat movie, dan respon pengguna di media sosial. Analisis sentimen pada movie review memiliki potensi untuk memberikan wawasan berharga kepada industri film, produser, dan penonton,

dengan membantu mereka dalam memahami bagaimana respon penonton terhadap suatu movie [2].

Salah satu dataset yang sangat berharga untuk analisis sentimen pada movie review adalah dataset IMDb. Internet Movie Database (IMDb) adalah database film terbesar dan terlengkap, yang menawarkan database ekstensif mengenai film, acara TV, dan informasi pemeran [3]. Namun, melakukan analisis sentimen movie review tidak dapat dikatakan mudah.

Dataset IMDb mencakup tantangan yang perlu diatasi dalam konteks analisis sentimen. Movie review sering mengandung bahasa yang kompleks, kosakata yang bervariasi, kalimat yang tidak terstruktur, dan lain sebagainya [4]. Tantangan ini juga mencakup penggunaan sumber daya dan waktu yang dibutuhkan dalam melakukan pemrosesan teks. Oleh karena itu, diperlukan pendekatan analisis sentimen yang efektif.

Salah satu metode yang dapat digunakan dalam penelitian ini adalah algoritma klasifikasi Naive Bayes. Algoritma Naive Bayes telah terbukti efektif dalam banyak kasus analisis sentimen karena sederhana, cepat, dan mampu memberikan hasil yang baik [14]. Namun, penerapannya dalam konteks movie review pada dataset IMDb memerlukan penilaian yang lebih lanjut.

Pada penelitian ini algoritma Naïve Bayes yang digunakan ialah Multinomial Naïve Bayes, algoritma ini cocok untuk data diskrit seperti teks karena bekerja dengan asumsi bahwa setiap atribut atau kata dalam teks diambil dari distribusi multinomial [5]. Selain itu digunakan metode TF-IDF pada vectorization untuk mengukur pentingnya kata-kata dalam review, dan chi-squared untuk memilih fitur-fitur (kata-kata) yang paling relevan dalam klasifikasi sentimen.

Penelitian ini akan mengevaluasi kinerja algoritma klasifikasi Multinomial Naïve Bayes menggunakan metrik evaluasi yang terdiri dari accuracy, precision, recall, dan F-1 score dalam mengatasi tantangan-tantangan yang mungkin muncul dalam analisis sentimen movie review.

Dengan melakukan penelitian ini, diharapkan akan ada kontribusi berharga dalam pemahaman analisis sentimen dalam konteks movie review pada dataset IMDb, seperti tantangan yang ada dan fitur yang relevan pada analisis sentimen, serta pemahaman dalam penggunaan algoritma klasifikasi Multinomial Naïve Bayes secara efektif.

## II. KAJIAN TEORI

Penelitian mengenai klasifikasi movie review sangat erat kaitannya dengan text mining yang menggunakan teks sebagai masukan data. Text mining atau penambangan data teks merupakan proses untuk menemukan informasi dengan menganalisis data tekstual. Proses mengacu pada cara pengambilan informasi berdasarkan pola yang ada pada teks [6].

Pada bidang sentimen analisis khususnya movie review, penelitian sangat penting dilakukan karena review film akan mempengaruhi keputusan pengguna untuk menonton atau tidak menonton suatu film [7]. Namun, menurut Pentheny, pengaruh ini tidak berlaku untuk semua tipe kepribadian orang.

Pada penelitian yang dilakukan oleh Sandesh Tripathi [8], dilakukan analisis sentimen movie review pada dataset IMDb dengan menggunakan metode klasifikasi Naïve Bayes, Logistic Regression, Random Forest, dan Decision Tree. Tahap pre-processing yang dilakukan diantaranya ialah menghapus tag HTML, dan normalisasi teks. Hal ini dilakukan karena sebagian besar movie review memiliki bahasa yang santai atau tidak formal, sehingga setiap kata yang memiliki makna sama perlu diubah menjadi bentuk dasarnya. Hasil dari penelitian menunjukkan bahwa Logistic Regression memiliki nilai accuracy terbaik yang kemudian diikuti oleh Naïve Bayes dan Random Forest.

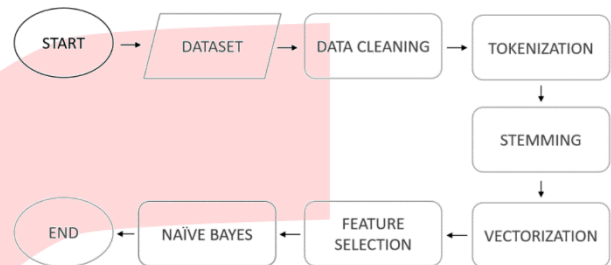
Atiqur Rahman [9] juga melakukan penelitian mengenai analisis sentimen movie review dengan menggunakan lima metode klasifikasi, dimana dua diantaranya ialah Bernoulli Naïve Bayes dan Multinomial Naïve Bayes. Pre-processing terdiri dari URL removal, bracket removal, tokenization, stop-words removal, case conversion, dan stemming. Penelitian tersebut menunjukkan bahwa Multinomial Naïve Bayes memiliki nilai accuracy, precision, dan F-1 score yang terbaik diantara lima metode tersebut. Sementara Bernoulli Naïve Bayes hanya mengalahkan Multinomial Naïve Bayes pada nilai recall saja.

Penelitian selanjutnya dilakukan oleh Ravinder Ahuja et al. [13] yang bertujuan untuk melihat pengaruh dari feature extraction pada analisis sentimen. Pada penelitian ini dibandingkan dua metode feature extraction, yaitu TF-IDF dan N-gram. Terlihat pada penelitian bahwa TF-IDF memiliki performa 3-4% lebih baik dalam analisis sentimen dibandingkan dengan N-gram, analisis dilakukan dengan menggunakan enam algoritma klasifikasi (Naïve Bayes, SVM, KNN, Decision Tree, Random Forest, Logistic Regression) dan metrik evaluasi berupa accuracy, precision, recall, dan F-1 score.

Muhamad Biki Hamzah [14] mengusulkan pendekatan analisis sentimen movie review pada dataset IMDb dengan menggunakan metode Multinomial Naïve Bayes, yang dibantu dengan chi-squared dan adaptive boosting. Chi-squared merupakan metode feature selection yang digunakan untuk mengurangi jumlah fitur yang digunakan, dimana pada kasus ini nilai chi-squared paling optimal ialah 2000. Accuracy yang dihasilkan pada penelitian ini dengan menggunakan algoritma Multinomial Naïve Bayes ialah 81.39%. Sedangkan dengan menggunakan metode chi-squared, algoritma memiliki nilai accuracy sebesar 85.37%.

## III. METODE

Metode atau alur sistem yang akan dibangun, terdiri dari beberapa proses mulai dari menentukan dan mengidentifikasi dataset yang digunakan yaitu dataset IMDb. Kemudian dilakukan pre-processing data yang terdiri dari proses data cleaning, tokenization, stemming, vectorization, dan feature selection. Setelah itu akan dilakukan proses training model menggunakan metode klasifikasi Naïve Bayes dalam proses klasifikasi teks pada dataset tersebut. Terakhir, akan dilakukan evaluasi untuk menguji nilai accuracy, precision, recall, dan F-1 score dari model klasifikasi yang dihasilkan. Selain itu, digunakan juga confusion matrix untuk mendukung metrik evaluasi yang digunakan.



GAMBAR 1  
(Alur sistem yang dibangun)

### A. Dataset

Pada penelitian sentimen analisis movie review ini digunakan dataset IMDb yang dapat ditemukan pada <https://www.kaggle.com/datasets/> [10]. Dataset yang digunakan ini berisi 50000 data, terdiri dari 25000 data positif dan 25000 data negatif. Dataset hanya berisikan review dan sentimen terhadap review tersebut, sehingga data dapat langsung diolah. Karena bentuk dataset yang non-linear, serta jumlah data yang tergolong banyak, performansi dari model Naïve Bayes dapat terpengaruh. Lama proses training dan evaluation data dari model juga akan terlihat.

### B. Data Cleaning

Proses ini dilakukan dengan mengkonversi data teks menjadi huruf kecil, menghapus tag HTML, menghapus tanda baca, dan menghapus karakter khusus seperti simbol dan emoji. Dengan melakukan langkah-langkah pembersihan data ini, data teks akan menjadi lebih standar, konsisten, dan hanya berfokus pada bagian yang penting. Hal ini tentunya dapat meningkatkan accuracy dan efektivitas algoritma klasifikasi yang digunakan, terutama untuk kasus seperti analisis sentimen.

### C. Tokenization

Tokenisasi adalah proses memecah teks menjadi unit-unit yang lebih kecil, disebut token. Token dapat berupa kata, frasa, atau simbol tertentu yang memiliki arti dalam bahasa tertentu. Tokenisasi penting dalam pemrosesan teks karena membantu mengubah teks menjadi format yang lebih mudah diolah oleh komputer.

### D. Stemming

Algoritma stemming yang digunakan ialah stemming porter. Algoritma ini digunakan untuk menghilangkan suffix dari kata-kata yang ada, dimana untuk kasus ini ialah kata dalam bahasa Inggris. Proses menghapus suffix ini

merupakan hal yang penting pada operasi pencarian informasi karena dengan begitu kata-kata tersebut dapat diurai menjadi bentuk kata dasarnya sehingga kualitas informasi yang dihasilkan akan meningkat [11].

#### E. Vectorization

Setelah tahap sebelumnya selesai, dilakukan proses pengambilan ciri-ciri (feature) dari data yang diolah. Vectorization ini bertujuan untuk memperkecil jumlah data, dan mengambil informasi yang terpenting dari data tersebut. Metode vectorization yang akan digunakan ialah Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF adalah metode yang berguna untuk menghitung bobot setiap kata yang umum digunakan. Nilai TF-IDF akan meningkat secara proporsional dengan tingkat kemunculan suatu term dalam suatu dokumen dan akan menurun dengan jumlah dokumen dalam korpus yang berisikan term tersebut [12].

#### F. Feature Selection

Penggunaan metode feature selection ini bertujuan untuk mengidentifikasi dan memilih fitur (kata-kata atau token dalam teks) yang paling informatif dan relevan untuk model yang dibangun [17]. Dengan mengurangi dimensi fitur, efisiensi model dapat ditingkatkan serta dapat menghindari masalah seperti overfitting. Metode seleksi fitur yang digunakan disini adalah chi-squared. Chi-squared merupakan metode statistik yang dapat digunakan untuk menentukan apakah terdapat hubungan signifikan antara dua variabel kategorikal [15].

#### G. Naïve Bayes

Setelah semua tahap pre-processing selesai, selanjutnya ialah mengimplementasikan algoritma klasifikasi Naïve Bayes untuk mulai melakukan proses klasifikasi pada dataset yang telah disiapkan. Naïve Bayes bekerja dengan menghitung probabilitas dalam membuat prediksi. Algoritma ini merupakan algoritma supervised learning, yang didasarkan pada teorema Bayes dan digunakan untuk memecahkan masalah klasifikasi. Teorema Bayes atau Hukum Bayes digunakan untuk menentukan probabilitas hipotesis berdasarkan pengetahuan sebelumnya [5]. Rumus dari Teorema Bayes ini ialah:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

- $P(A|B)$  merupakan probabilitas bersyarat A yang diberikan oleh B.
- $P(B|A)$  merupakan probabilitas bersyarat B yang diberikan oleh A.
- $P(A)$  merupakan probabilitas kejadian A.
- $P(B)$  merupakan probabilitas kejadian B.

Pada analisis sentimen, probabilitas review tertentu termasuk kedalam kelas positif atau negatif ditentukan berdasarkan kata-kata yang ada dalam review tersebut. Oleh karena itu, data teks sebelumnya harus diubah menjadi format numerik sehingga dapat dipahami oleh algoritma. Terdapat beberapa jenis algoritma Naïve Bayes yang dapat digunakan, untuk kasus ini algoritma yang digunakan ialah Multinomial Naïve Bayes. Algoritma ini cocok untuk data diskrit seperti teks karena bekerja dengan asumsi bahwa setiap atribut atau

kata dalam teks diambil dari distribusi multinomial [5]. Artinya, algoritma Multinomial Naïve Bayes menghitung frekuensi kemunculan kata-kata dalam setiap kelas sentimen dan menggunakannya untuk membuat prediksi. Algoritma ini cenderung cukup cepat dan efisien, bahkan untuk dataset besar seperti pada kasus ini, dan juga secara efektif dapat mengekstrak informasi dari frekuensi kata dalam teks. Ini dapat membantu mengidentifikasi kata-kata yang paling menggambarkan sentimen positif atau negatif.

Dalam mengukur performansi dari hasil penelitian yang dikerjakan, dilakukan proses evaluasi dengan menghitung nilai accuracy, f1-score, recall, dan precision terhadap kedua model klasifikasi.

#### H. Accuracy

Merupakan kebenaran model dalam melakukan klasifikasi data dengan benar. Accuracy mengukur tingkat kebenaran dari model dengan menghitung rasio prediksi yang benar (baik true positives dan true negatives) terhadap keseluruhan data dalam dataset. Berikut merupakan persamaan accuracy:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

#### I. Recall

Menggambarkan seberapa baik model dalam mengidentifikasi data positif pada dataset. Recall menghitung rasio true positives terhadap total jumlah prediksi positif sebenarnya (true positives dan false negatives). Berikut merupakan persamaan recall:

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

#### J. Precision

Precision berfokus pada accuracy dari prediksi positif yang dibuat oleh model. Precision mengukur rasio true positives terhadap total jumlah prediksi positif (true positives dan false positives). Berikut merupakan persamaan precision:

$$precision = \frac{TP}{TP + FP} \quad (4)$$

#### K. F-1 Score

Menggambarkan rata-rata harmonik antara recall dan precision. F1-Score memberikan keseimbangan antara kedua metrik ini dan akan sangat berguna dalam menghadapi kelas yang tidak seimbang. Berikut adalah persamaannya:

$$f1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (5)$$

#### L. Confusion Matrix

Digunakan untuk mengevaluasi kinerja model klasifikasi dengan merangkum prediksinya dan membandingkannya dengan hasil actual [16]. Confusion matrix terdiri dari empat nilai atau istilah yang menggambarkan hasil representasi proses klasifikasi model, yaitu:

- True Positives (TP): jumlah data positif yang diprediksi benar oleh model sebagai positif.
- True Negatives (TN): jumlah data negatif yang diprediksi benar oleh model sebagai negatif.
- False Positives (FP): jumlah data negatif yang salah diprediksi oleh model sebagai positif.

- False Negatives (FN): jumlah data positif yang salah diprediksi oleh model sebagai negatif.

#### IV. HASIL DAN PEMBAHASAN

##### A. Hasil Pengujian

Dalam pengujian ini dilakukan analisis sentimen pada dataset IMDb yang terdiri dari 50.000 data dengan menggunakan Multinomial Naïve Bayes, serta Bernoulli Naïve Bayes sebagai pembanding. Pada pre-processing data tahap vectorization, parameter 'max\_features' dari TF-IDF diberikan nilai 50000 untuk menentukan jumlah fitur yang diambil berdasarkan frekuensinya. Sedangkan pada tahap feature selection, parameter 'k' atau 'num\_features' dari chi-squared diberikan nilai 15000 untuk menentukan jumlah fitur terbaik yang akan diambil.

Hasil pengujian menunjukkan bahwa accuracy model Multinomial Naïve Bayes mencapai 87.71%, dengan nilai precision, recall, dan f1-score yang juga mencapai 0.87. Dengan hasil yang baik tersebut, proses training model sangat cepat dan tidak intensif. Berikut merupakan confusion matrix dari model Multinomial Naïve Bayes tersebut:

		Actual	
		TP	FP
Predicted	TP	4386	575
	FN	654	4385

GAMBAR 2  
(Confusion Matrix Multinomial Naïve Bayes)

Dapat dilihat bahwa model melakukan 4386 prediksi benar terhadap sentimen positif dan 4385 prediksi benar terhadap sentimen negatif. Namun, terdapat 575 prediksi positif palsu (kasus sentimen negatif yang salah diklasifikasikan sebagai positif) dan 654 prediksi negatif palsu (kasus sentimen positif yang salah diklasifikasikan sebagai negatif).

Sementara hasil pengujian Bernoulli Naïve Bayes menunjukkan bahwa model dapat mencapai 87.41% accuracy, dengan nilai precision, recall, dan f1-score yang juga mencapai 0.87. Hasil yang didapatkan tersebut sangat mirip dengan hasil Multinomial Naïve Bayes, dimana hanya berbeda 0.3% saja. Proses training dan evaluasi model juga sama dan tidak ada perbedaan yang signifikan. Berikut merupakan confusion matrix dari model Bernoulli Naïve Bayes tersebut:

		Actual	
		TP	FP
Predicted	TP	4364	597
	FN	662	4377

GAMBAR 3  
(Confusion Matrix Bernoulli Naïve Bayes)

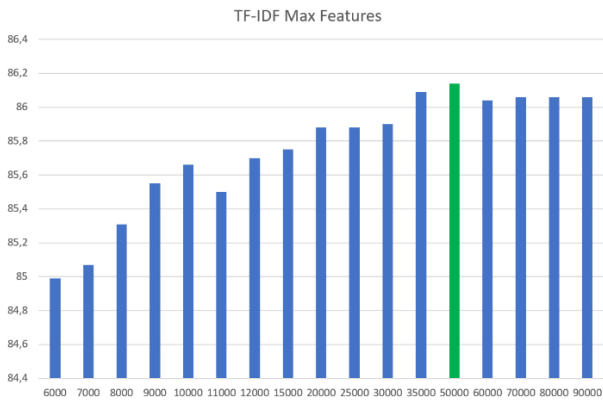
Terlihat bahwa model melakukan 4364 prediksi benar terhadap sentimen positif dan 4377 prediksi benar terhadap sentimen negatif. Namun, terdapat 597 prediksi positif palsu (kasus sentimen negatif yang salah diklasifikasikan sebagai positif) dan 662 prediksi negatif palsu (kasus sentimen positif yang salah diklasifikasikan sebagai negatif).

Berdasarkan hasil dari pengujian ini, dapat dilihat bahwa kedua model algoritma klasifikasi Naïve Bayes sangat baik dalam melakukan klasifikasi movie review pada dataset IMDb. Namun, algoritma Multinomial Naïve Bayes sedikit lebih baik dibandingkan dengan Bernoulli Naïve Bayes.

##### B. Analisis Hasil Pengujian

Pada penelitian yang dilakukan, pemilihan nilai optimal untuk parameter 'max\_features' dalam vektorisasi TF-IDF bergantung pada beberapa faktor, seperti dataset, sumber daya komputasi yang tersedia, dan tujuan analisis. Dataset yang digunakan terdiri dari 50.000 data, sehingga nilai 'max\_features' akan sangat berpengaruh karena ukuran dataset yang mungkin tergolong besar. Jika dataset relatif kecil, dapat digunakan nilai 'max\_features' yang besar untuk mengambil lebih banyak fitur tanpa adanya beban komputasi. Namun, untuk dataset yang sangat besar, sebaiknya digunakan nilai yang lebih kecil untuk mengurangi kompleksitas komputasi. Jika memori atau daya pemrosesan yang dimiliki terbatas, menggunakan nilai yang besar untuk 'max\_features' mungkin akan sangat sulit karena waktu pemrosesan yang lama. Nilai yang besar juga dapat memasukkan noise dari fitur-fitur yang kurang relevan. Sementara itu, nilai kecil untuk 'max\_features' dapat menyebabkan hilangnya informasi jika fitur-fitur penting tidak dimasukkan pada TF-IDF. Sehingga pada penelitian, selain perlu menyeimbangkan pengambilan fitur yang banyak dengan batasan perangkat keras, menjaga fitur-fitur yang paling informatif sambil menghindari noise juga perlu diperhatikan. Berikut merupakan beberapa hasil uji coba yang dilakukan untuk menemukan nilai 'max\_features' yang optimal:

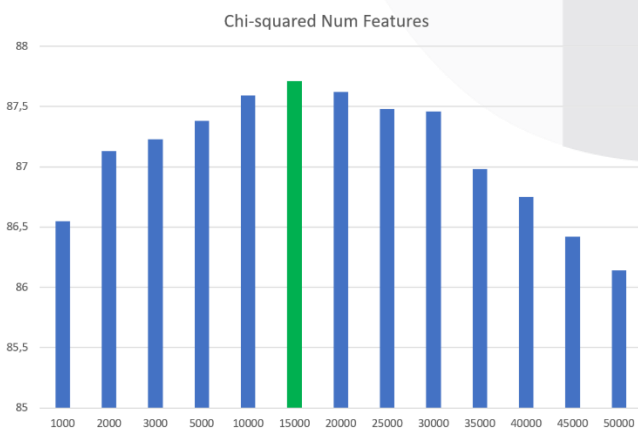




GAMBAR 4  
(Hasil accuracy model berdasarkan nilai TF-IDF)

Dapat dilihat pada gambar diatas, percobaan dilakukan dengan menggunakan nilai 'max\_features' sekecil 6000 yang menghasilkan accuracy bernilai 84.99%, sampai dengan nilai 'max\_features' sebesar 90000 yang menghasilkan accuracy bernilai 86.06%. Dari sekian banyak percobaan yang dilakukan, 'max\_features' sebesar 50000 merupakan nilai yang paling optimal karena dapat menghasilkan nilai accuracy paling tinggi terhadap model.

Nilai terbaik pada percobaan tersebut digunakan untuk mencari nilai optimal pada parameter 'k' atau 'num\_features' dalam melakukan feature selection dengan metode chi-squared. Penentuan nilai optimal pada kasus ini juga sama seperti sebelumnya, yaitu tergantung pada dataset, serta tujuan dan metrik evaluasi yang digunakan. Oleh karena itu, dilakukan eksperimen dengan berbagai nilai 'num\_features' dan diukur pengaruhnya pada performa model berdasarkan nilai akurasi, presisi, recall, dan F1-score yang dihasilkan. Nilai 'num\_features' yang lebih besar akan mengambil lebih banyak fitur dalam model, yang dapat meningkatkan kemampuan model pada proses training. Namun, nilai 'num\_features' yang terlalu besar juga dapat menyebabkan adanya noise dan membuat model lebih kompleks. Sehingga perlu ditemukan keseimbangan yang baik antara meningkatkan informasi yang relevan dan mengurangi kompleksitas yang tidak perlu. Berikut merupakan hasil uji coba yang dilakukan untuk menemukan nilai 'num\_features' yang optimal, dengan nilai 'max\_features' sebesar 50000:



GAMBAR 5  
(Hasil accuracy model berdasarkan nilai chi-squared)

Gambar tersebut menunjukkan hasil evaluasi model dari setiap percobaan yang dilakukan berdasarkan parameter yang digunakan, dimulai dari 'num\_features'=1000 yang menghasilkan accuracy bernilai 86.55%, sampai dengan 'num\_features'=50000 yang menghasilkan accuracy bernilai 86.15%. Terlihat bahwa 'num\_features' dengan nilai 15000 memiliki accuracy yang terbaik, yaitu sebesar 87.71%. Sehingga nilai optimal yang didapatkan berdasarkan hasil evaluasi model ialah 'num\_features'=15000 dengan 'max\_features'=50000.

TABEL 1  
(Beberapa contoh fitur yang paling relevan)

Feature	Score
jamon	0.6901
bradfield	0.6900
mindset	0.6900
cantillana	0.6899
burp	0.6899
weatherli	0.6899
uncompel	0.6898
kureishi	0.6897
oneil	0.6896
time310	0.6896
failedi	0.6896
softspoken	0.6895
rabin	0.6893
nearsight	0.6892
attach	0.6888

Dapat dilihat pada tabel diatas fitur-fitur yang termasuk paling relevan, ditandai dengan nilai skor tinggi. Pengaruh fitur yang paling relevan terhadap model dapat sangat besar dalam analisis data dan pemodelan. Fitur ini merupakan fitur-fitur yang memberikan kontribusi besar dalam membuat model memahami dan memprediksi data dengan akurat, serta cenderung memiliki hubungan yang kuat dengan variabel target. Oleh karena itu, kinerja model terkait accuracy, precision, recall, dan F-1 score dapat meningkat.

Menentukan fitur yang tepat dapat menjadi tantangan dalam melakukan analisis sentimen movie review. Hal ini dapat disebabkan oleh penggunaan bahasa yang tidak formal dan ekspresif, dimana reviewer juga dapat menggunakan bahasa metaforis, idiom, atau kata-kata khusus untuk mengungkapkan perasaan mereka. Beberapa movie review mungkin terdiri dari kalimat yang tidak lengkap atau tidak terstruktur dengan baik, termasuk pemakaian tata bahasa yang tidak standar, seperti ejaan alternatif, singkatan, atau slang. Hal ini membuat proses feature selection yang tepat menjadi lebih rumit, karena kata-kata tersebut mungkin tidak memiliki makna yang jelas atau umum.

Setiap metode pre-processing yang dilakukan pada penelitian ini bertujuan untuk mengatasi dan meminimalisir tantangan ini, dimulai dari proses data cleaning, tokenization, stemming, vectorization, hingga feature selection. Terkait

proses stemming, hal ini dilakukan karena penggunaan kata-kata dalam bentuk dasar akan membantu mengurangi kompleksitas dataset. Pada dataset IMDb ini, terdapat banyak kata-kata yang memiliki makna yang sama tetapi berbeda dalam bentuk atau konjugasi sehingga dapat mempengaruhi hasil analisis. Dengan melakukan stemming, model akan lebih mudah mengidentifikasi kesamaan makna dalam review yang menggunakan variasi kata berbeda. Namun, proses ini dapat menghasilkan kata-kata yang mungkin tidak valid dalam penggunaan bahasa normal, contohnya seperti beberapa fitur diatas. Hal tersebut mungkin sulit untuk dipahami, tetapi akan tetap berguna dalam pemrosesan teks untuk melakukan klasifikasi. Ini adalah salah satu langkah yang membantu meminimalisir salah satu tantangan yang dihadapi dalam analisis sentimen ulasan film, yaitu variasi kata dalam bahasa yang tidak formal dan ekspresif.

## V. KESIMPULAN

Berdasarkan hasil pengujian dan analisis yang telah dilakukan terhadap metode klasifikasi Naïve Bayes dalam analisis sentimen movie review pada dataset IMDb, metode Naïve Bayes yang digunakan tersebut terbukti efektif. Kedua metode Naïve Bayes yang digunakan, yaitu Multinomial Naive Bayes dan Bernoulli Naive Bayes memiliki tingkat accuracy, precision, recall, dan F1-score yang tinggi, menunjukkan kemampuan yang baik dalam mengklasifikasikan sentimen positif dan negatif pada movie review. Metode ini merupakan metode yang cepat dan sederhana dalam melakukan klasifikasi, sehingga efisien dalam waktu komputasi.

Tantangan pada dataset IMDb dapat berupa ukuran dataset yang relatif besar, pemilihan fitur yang kurang tepat, penggunaan bahasa informal, metaforis, idiom, atau kata-kata khusus, serta kalimat yang tidak lengkap dan terstruktur. Dalam mengatasi dan meminimalisir tantangan, dilakukan tahap pre-processing yang terdiri dari data cleaning, tokenization, stemming, vectorization (TF-IDF), dan feature selection (chi-squared) dengan mencari nilai parameter optimal. Nilai yang paling optimal untuk TF-IDF ialah 'max\_features'=50000 yang menghasilkan nilai accuracy sebesar 86.14%. Sedangkan nilai paling optimal untuk chi-squared ialah 'num\_features'=15000, dengan nilai accuracy sebesar 87.71%. Beberapa contoh fitur yang paling relevan ialah uncompele (uncompelling) dan attach (attachment), yang artinya fitur-fitur tersebut cenderung memiliki hubungan yang erat terhadap nilai sentimen.

Dalam melanjutkan dan mengembangkan analisis sentimen movie review pada IMDb, satu saran yang dapat diberikan ialah dengan peningkatan model klasifikasi. Peningkatan model klasifikasi dilakukan dengan eksplorasi berbagai model klasifikasi lainnya seperti Decision Trees, Random Forests, Neural Networks, yang dapat memberikan pengetahuan lebih lanjut tentang performa dan kemampuan model dalam analisis sentiment movie review.

## REFERENSI

- [1] Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167.
- [2] Gupta, S. Movie Review Sentiment Analysis Using Machine Learning. URL: <https://www.enjoyalgorithms.com/blog/sentiment-analysis-using-machine-learning>. (accessed: 01.09.2023).
- [3] Fisher, S. 2022. Apa itu IMDb?. URL: [https://www-lifewire-com.translate.goog/internet-movie-database-3482288?\\_x\\_tr\\_sl=en&\\_x\\_tr\\_tl=id&\\_x\\_tr\\_hl=id&\\_x\\_tr\\_pto=tc](https://www-lifewire-com.translate.goog/internet-movie-database-3482288?_x_tr_sl=en&_x_tr_tl=id&_x_tr_hl=id&_x_tr_pto=tc). (accessed: 01.09.2023).
- [4] Sharma, H., Pangaonkar, S., Gunjan, R., dan Rokade, P. 2023. Sentimental Analysis of Movie Reviews Using Machine Learning. *ITM Web of Conferences*, vol. 53.
- [5] Naïve Bayes Classifier Algorithm. URL: <https://www.javatpoint.com/machine-learning-naive-bayes-classifier>. (accessed: 01.09.2023).
- [6] Tan, A. 1999. Text Mining: The state of the art and the challenges. *Proc. PAKDD 1999 Work. Knowledge Discovery from Adv. Databases*, vol. 8, pp. 65–70.
- [7] Pentheny, J. R. 2015. *The Influence of Movie Reviews on Consumers*. University of New Hampshire.
- [8] Tripathi, S., Mehrotra, R., Bansal, V., dan Upadhyay, S. 2020. Analyzing Sentiment using IMDb Dataset. *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*.
- [9] Rahman, A., dan Hossen, S. 2019. Sentiment Analysis on Movie Review Data Using Machine Learning Approach. *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*.
- [10] Lakshmipathi, N. 2019. IMDb Datasets of 50k Movie Reviews. URL: <https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-moviereviews>. (accessed: 04.05.2021).
- [11] Indriyono, B. V., Utami, E., dan Sunyoto, A. 2015. Pemanfaatan Algoritma Porter Stemmer Untuk Bahasa Indonesia Dalam Proses Klasifikasi Jenis Buku. *Jurnal Buana Informatika*, vol. 6, no. 4, 301-310.
- [12] Kothari, M. 2020. Feature Extraction Techniques - NLP. URL: <https://www.geeksforgeeks.org/feature-extraction-techniques-nlp/>. (accessed: 22.06.2021).
- [13] Ahuja, R., Chug, A., Kohli, S., Gupta, S., dan Ahuja, P. 2019. The Impact of Features Extraction on the Sentiment Analysis. *Procedia Computer Science*, 152, 341–348.
- [14] Hamzah, B. 2021. Classification of Movie Review Sentiment Analysis Using Chi-Square and Multinomial Naïve Bayes with Adaptive Boosting. *Journal of Advances in Information Systems and Technology* 3(1), 67-74.
- [15] Gupta, A. 2023. Chi-Square Test for Feature Selection – Mathematical Explanation. URL: <https://www.geeksforgeeks.org/chi-square-test-for-feature-selection-mathematical-explanation/>. (accessed: 01.09.2023).
- [16] Solichin, A. 2017. Mengukur Kinerja Algoritma Klasifikasi dengan Confusion Matrix. URL: <https://achmatim.net/2017/03/19/mengukurkinerja-algoritma-klasifikasi-dengan-confusion-matrix>. (accessed: 07.05.2021).
- [17] Azevedo, N. 2021. Data Preprocessing: 6 Techniques to Clean Data. URL: <https://www.scalablepath.com/data-science/data-preprocessing-phase>. (accessed: 13.06.2023).