KLASIFIKASI SENTIMEN ULASAN FILM IMDB MENGGUNAKAN METODE N-GRAM DAN LOGISTIC REGRESSION

Michael Angelo Jordan [1] Viny Christanti Mawardi, S.Kom., M.Kom. [2]

[1) 2)] Teknik Informatika Universitas Tarumanagara
Letjen S. Parman St No.1, RT.6/RW.16, Tomang, Grogol petamburan, Kota Jakarta Barat, Jakarta 11440
email : michael.535220246@stu.untar.ac.id [1)], vinym@fti.untar.ac.id [2)]

ABSTRAK

Analisis sentimen merupakan salah satu penerapan Natural Language Processing (NLP) yang digunakan untuk mengidentifikasi opini atau emosi seseorang terhadap suatu objek, produk, atau layanan. Dalam penelitian ini, dilakukan analisis sentimen terhadap ulasan film pada dataset IMDb dengan tujuan untuk mengklasifikasikan ulasan menjadi sentimen positif atau negatif. Metode yang digunakan adalah N-Gram sebagai teknik ekstraksi fitur teks dan Logistic Regression sebagai algoritma klasifikasi. Proses diawali dengan pra-pemrosesan teks yang mencakup case folding, penghapusan stopwords, dan lemmatization menggunakan WordNetLemmatizer. Selanjutnya, data direpresentasikan menggunakan TF-IDF (Term Frequency–Inverse Document Frequency) dengan kombinasi N-Gram (1–3) untuk menangkap konteks kata berurutan. Hasil pengujian menunjukkan bahwa model yang dihasilkan memiliki tingkat akurasi sebesar 83%, dengan performa yang baik dalam mendeteksi sentimen positif maupun negatif. Meskipun demikian, model masih memiliki keterbatasan dalam memahami konteks kalimat negasi seperti "not bad" atau "no good".

Key words

Analisis Sentimen, NLP, Logistic Regression, N-Gram, TF-IDF, IMDb

1. Pendahuluan

Perkembangan teknologi informasi dan komunikasi telah menghasilkan jumlah data teks yang sangat besar di internet. Data teks ini sering kali berisi opini dan emosi pengguna, yang dapat dimanfaatkan untuk mengetahui persepsi masyarakat terhadap suatu produk atau layanan. Salah satu bentuk pemanfaatan analisis teks tersebut adalah *Sentiment Analysis* atau analisis sentimen, yaitu proses mengidentifikasi polaritas opini dalam suatu teks apakah bersifat positif, negatif, atau netral.

IMDb (*Internet Movie Database*) merupakan salah satu sumber data populer yang berisi jutaan ulasan pengguna terhadap berbagai film. Analisis sentimen pada data IMDb dapat membantu studio film, perusahaan, dan pengguna umum dalam memahami tanggapan publik secara otomatis.

Penelitian ini bertujuan untuk membangun model analisis sentimen menggunakan metode *N-Gram* dan algoritma *Logistic Regression*. Pendekatan ini dipilih karena bersifat sederhana namun mampu memberikan hasil yang cukup baik dalam klasifikasi teks. Hasil penelitian ini diharapkan dapat memberikan pemahaman mengenai penerapan metode *N-Gram* dan *Logistic Regression* pada tugas klasifikasi sentimen berbasis teks.

1.1 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana menerapkan metode *N-Gram* dan *Logistic Regression* untuk melakukan klasifikasi sentimen terhadap ulasan film IMDb?
2. Sejauh mana tingkat akurasi model dalam mengenali sentimen positif dan negatif pada data ulasan film IMDb?
3. Faktor apa saja yang mempengaruhi kesalahan prediksi sentimen, seperti pada kasus penggunaan negasi atau frasa ambigu?

1.2 Tujuan Penelitian

Adapun tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

1. Membangun model klasifikasi sentimen berbasis *Machine Learning* dengan menggunakan kombinasi metode *N-Gram* dan *Logistic Regression*.
2. Menganalisis performa model menggunakan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score.
3. Mengidentifikasi kelemahan model dalam mengenali pola kata tertentu yang menyebabkan kesalahan klasifikasi, serta memberikan rekomendasi untuk penelitian lanjutan.

2. Metode Penelitian

2.1. Dataset

Dataset yang digunakan adalah IMDb Movie Reviews Dataset yang diambil dari pustaka NLTK (Natural Language Toolkit) serta diperluas dengan data tambahan dari Kaggle. Dataset ini berisi ulasan film

berbahasa Inggris yang telah diberi label sentimen positif dan negatif.

Secara umum, dataset ini dipilih karena memiliki label sentimen yang jelas, bersih, dan terverifikasi, serta sering digunakan dalam berbagai penelitian analisis sentimen berbasis teks. Penjelasan lebih rinci mengenai karakteristik dan alasan pemilihan dataset dibahas pada subbab 3.1 Data set.

2.2. Pra-pemrosesan Teks

Tahapan pra-pemrosesan dilakukan agar teks menjadi seragam dan siap untuk proses pelatihan model. Langkah-langkah yang diterapkan meliputi:

1. Case folding – mengubah seluruh huruf menjadi huruf kecil.
2. Cleaning – menghapus tanda baca, angka, dan karakter non-alfabet.
3. Stopword removal – menghapus kata umum seperti “the”, “and”, “is”.
4. Lemmatization – mengubah kata ke bentuk dasarnya menggunakan WordNetLemmatizer.

2.3. Ekstraksi Fitur Menggunakan N-Gram dan TF-IDF

Setiap dokumen diubah menjadi representasi numerik menggunakan metode *Term Frequency–Inverse Document Frequency (TF-IDF)* yang dikombinasikan dengan *N-Gram* (1–3).

Rumus perhitungan *TF-IDF* adalah sebagai berikut:

$$TF - IDF(t,d) = TF(t,d) \times log\ log\ \frac{N}{DF(t)}$$

dengan:

1. $TF(t,d)$= frekuensi kemunculan term *t* dalam dokumen *d*
2. $DF(t)$= jumlah dokumen yang mengandung term *t*
3. N= total jumlah dokumen

Pendekatan *N-Gram* digunakan untuk menangkap konteks berurutan antar kata, misalnya:

1. Unigram: good, bad
2. Bigram: not good, very bad
3. Trigram: not at all bad

Dengan demikian, representasi teks tidak hanya bergantung pada kata tunggal, tetapi juga kombinasi kata yang memiliki makna emosional lebih spesifik.

2.4. Model Klasifikasi Logistic Regression

Kinerja model diuji menggunakan metrik Akurasi, Presisi, Recall, dan F1-score dengan rumus:

$$P(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n}}$$

Keterangan:

1. $P(y = 1 \mid x)$= probabilitas bahwa ulasan bersentimen positif
2. e: Bilangan exponensial(≈ 2.718).
3. β_0= bias/intercept
4. β_i= bobot dari fitur ke-i
5. x_i= nilai fitur TF-IDF untuk kata ke-i

Fungsi ini menghasilkan nilai antara 0 dan 1. Jika $P(y = 1 \mid x) > 0.5$, maka model memprediksi positif, sebaliknya negatif.

Fungsi Kerugian (Loss Function)

Proses pelatihan model dilakukan dengan meminimalkan kesalahan prediksi menggunakan Binary Cross-Entropy Loss:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i\ log(\hat{y}_i) + (1 - y_i)\ log\ (1 - \hat{y}_i)]$$

Penjelasan:

1. L: nilai kerugian rata-rata.
2. N: jumlah total data latih.
3. y_i: label sebenarnya dari data ke-i (0 = negatif, 1 = positif).
4. $\hat{y}_i$: probabilitas hasil prediksi oleh model.
5. Logaritma digunakan untuk memperbesar penalti terhadap kesalahan prediksi yang besar.

Model akan menyesuaikan bobot (β_i) secara iteratif menggunakan algoritma optimasi (misalnya *liblinear solver*) hingga fungsi kerugian L minimum.

2.5. Evaluasi Model

Evaluasi model dilakukan menggunakan metode K-Fold Cross Validation, yaitu membagi dataset menjadi k bagian (fold). Dalam penelitian ini digunakan k = 5.

Setiap iterasi:

1. 4 bagian digunakan untuk training
2. 1 bagian digunakan untuk testing
3. Proses diulang sebanyak 5 kali, lalu hasil akurasi dirata-ratakan.

Rumus Evaluasi K-Fold:

$$Accuracy_{CV} = \frac{1}{k}\sum_{i=1}^{k} Accuracy_i$$

Penjelasan:

1. k: jumlah lipatan (fold).
2. $Accuracy_i$: akurasi hasil pengujian pada lipatan ke-i.
3. $Accuracy_{CV}$: rata-rata akurasi keseluruhan model.

Metode ini memastikan hasil evaluasi tidak bias terhadap pembagian data tertentu. Selain itu, performa model juga diukur menggunakan metrik-metrik berikut:

1. Akurasi:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

Mengukur persentase prediksi yang benar dari seluruh data uji.

2. Presisi

$$Presisi = \frac{TP}{TP + FP}$$

Menunjukkan seberapa banyak hasil positif yang benar dari semua prediksi positif.

3. Recall

$$Recall = \frac{TP}{TP + FN}$$

Mengukur kemampuan model dalam menemukan semua data positif yang benar.

4. F1-Score

$$F1 - Score = 2 \times \frac{Presisi \times Recall}{Presisi + Recall}$$

F1-score digunakan untuk menilai keseimbangan antara presisi dan recall.

Penjelasan:

1. TP (True Positive):
   Jumlah data ulasan yang sebenarnya positif dan diprediksi positif oleh model. Contoh: "This movie is fantastic!" diprediksi positif, label aslinya positif.
2. TN (True Negative):
   Jumlah data ulasan yang sebenarnya negatif dan diprediksi negatif oleh model. Contoh: "This movie was terrible." diprediksi negatif, label aslinya negatif.
3. FP (False Positive):
   Jumlah data ulasan yang sebenarnya negatif tetapi diprediksi positif. Contoh: "Not good at all." diprediksi positif, padahal negatif.
4. FN (False Negative):
   Jumlah data ulasan yang sebenarnya positif tetapi diprediksi negatif. Contoh: "Not bad." diprediksi negatif, padahal positif.

2.6. Alur Penelitian

Untuk mencapai tujuan penelitian, dilakukan serangkaian tahapan sistematis mulai dari pengumpulan data hingga evaluasi hasil model. Setiap tahapan saling berhubungan dan berperan penting dalam membangun sistem klasifikasi sentimen yang akurat. Tahapan awal berfokus pada persiapan data teks agar dapat diproses oleh komputer, dilanjutkan dengan transformasi data menjadi representasi numerik, pelatihan model, serta pengujian dan evaluasi kinerja model. Secara keseluruhan, alur penelitian ini menggambarkan proses end-to-end dari *Natural Language Processing (NLP)* dalam mengklasifikasikan sentimen ulasan film pada dataset IMDb.

Adapun langkah-langkah penelitian yang dilakukan adalah sebagai berikut:

1. Dataset yang digunakan dalam penelitian ini adalah IMDb Movie Reviews Dataset yang tersedia di pustaka NLTK (Natural Language Toolkit) serta diperluas dengan data tambahan dari Kaggle. Total dataset yang digunakan berjumlah 100.000 dokumen teks, terdiri atas 50.000 ulasan positif dan 50.000 ulasan negatif.
2. Melakukan pra-pemrosesan teks (case folding, cleaning, stopword removal, lemmatization). Tujuannya untuk membersihkan dan menstandarkan data teks agar lebih mudah dianalisis oleh algoritma.
3. Mengubah teks menjadi vektor numerik menggunakan TF-IDF dan N-Gram. Proses ini mengubah teks menjadi representasi matematis yang mencerminkan pentingnya setiap kata dan urutan katanya.
4. Melatih model Logistic Regression menggunakan data latih. Model akan belajar dari pola yang terdapat pada data latih untuk membedakan ulasan positif dan negatif.
5. Melakukan evaluasi model dengan K-Fold Cross Validation dan menghitung akurasi rata-rata. Evaluasi ini dilakukan untuk memastikan bahwa model memiliki performa yang stabil di seluruh dataset.
6. Menguji model dengan ulasan baru untuk memverifikasi kemampuan prediksi. Tahap ini berfungsi untuk menguji kehandalan model dalam menghadapi data yang belum pernah dilihat sebelumnya.

3. Hasil dan Pembahasan

Setelah seluruh tahap pra-pemrosesan dan ekstraksi fitur selesai dilakukan, data teks yang telah direpresentasikan dengan metode TF-IDF dan N-Gram digunakan untuk melatih model Logistic Regression. Proses pelatihan dilakukan menggunakan parameter max_iter=3000 dan solver 'liblinear' untuk memastikan konvergensi model yang optimal.

Model dilatih dengan rasio pembagian data 70% untuk training set dan 30% untuk testing set. Selain itu, untuk memastikan kestabilan performa, dilakukan juga evaluasi menggunakan metode 5-Fold Cross Validation.

Hasil pelatihan menunjukkan bahwa model mampu mempelajari pola kata yang berhubungan dengan sentimen, seperti kata "good", "amazing", dan "excellent" untuk ulasan positif, serta "bad", "boring", dan "waste" untuk ulasan negatif. Hal ini menunjukkan bahwa kombinasi TF-IDF dan N-Gram berhasil membantu model memahami konteks lokal antar kata dalam teks.

3.1. Dataset

Dataset yang digunakan dalam penelitian ini adalah IMDb Movie Reviews Dataset yang tersedia dalam pustaka NLTK (Natural Language Toolkit). Dataset ini merupakan kumpulan ulasan film berbahasa Inggris dari situs Internet Movie Database (IMDb) yang telah diberi label sentimen secara manual. Dataset ini pertama kali diperkenalkan oleh Maas et al. (2011) dan menjadi salah satu dataset standar dalam penelitian analisis sentimen. Selain itu ada 50.000 data tambahan dari dataset yang bersumber dari Kaggle untuk menambahkan data training sehingga dapat menambahkan tingkat akurasi sistem. Total data yang digunakan dalam penelitian ini berjumlah 100.000 dokumen teks, yang terdiri dari:

1. 50.000 ulasan positif, yaitu ulasan dengan nada emosional yang menggambarkan kepuasan terhadap film,

2. 50.000 ulasan negatif, yaitu ulasan dengan nada emosional yang menunjukkan kekecewaan terhadap film.

Setiap dokumen merupakan satu paragraf ulasan film dengan panjang bervariasi, rata-rata antara 50 hingga 300 kata. Dataset ini dipilih karena:

1. Sudah memiliki label sentimen yang bersih dan terverifikasi, sehingga cocok digunakan untuk pelatihan model klasifikasi.
2. Mewakili beragam gaya bahasa pengguna internet, termasuk penggunaan idiom, negasi, dan sarkasme ringan.
3. Banyak digunakan pada penelitian terdahulu, sehingga hasil eksperimen dapat dibandingkan dengan penelitian sebelumnya secara objektif.

Struktur data dalam korpus NLTK ini tersusun sebagai berikut:

Tabel 1.1 Isi output deskripsi dari dataset

Kolom	Deskripsi
review	Teks ulasan film berformat string.
label	Kategori sentimen dari ulasan, terdiri dari pos (positif) dan neg (negatif).

Contoh data yang diambil dari korpus NLTK ditunjukkan pada Tabel 1.2 berikut:

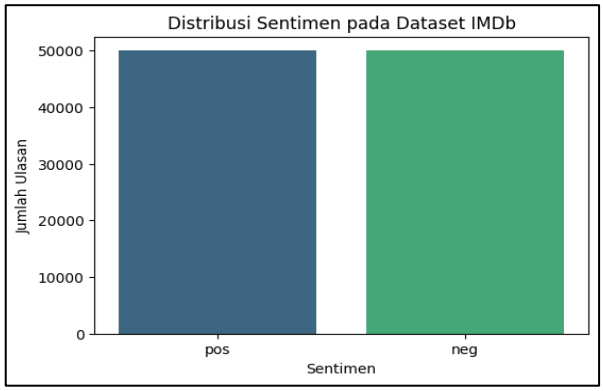
Tabel 1. 2 Contoh data yang di ambil dari korpus NLTK

No	**Review (potongan teks)**	**Label**
1	"An amazing performance by the cast and a touching storyline."	Positive
2	"This movie was a total waste of time, very disappointing."	Negative
3	"Good visuals but the plot was dull and predictable."	Negative
4	"A masterpiece that keeps you thinking long after it ends."	Positive

Dataset ini kemudian diacak (shuffled) untuk memastikan distribusi data acak dan seimbang antara dua kelas. Langkah ini penting untuk menghindari bias selama proses pelatihan model. Selanjutnya, dataset dibagi menjadi dua bagian:

1. Data latih (train set): 70% dari total dataset, digunakan untuk melatih model Logistic Regression.
2. Data uji (test set): 30% dari total dataset, digunakan untuk mengevaluasi performa model secara objektif.

Pembagian dataset dapat di lihat secara visual, distribusi kelas antar data positive dengan data Negative ditunjukkan dalam bentuk bar chart pada Gambar 1.1.



Gambar 1.1 Distribusi Data Sentime IMDb

Distribusi data yang seimbang ini menjadi keunggulan dataset IMDb, karena menghindarkan model dari bias terhadap salah satu kelas (class imbalance). Selain itu, dataset juga mengandung variasi struktur kalimat dan kata-kata ekspresif seperti *"amazing"*, *"boring"*, *"waste"*, dan *"excellent"*, yang sangat membantu model dalam membedakan nuansa emosional setiap ulasan.

3.2. Hasil Pra-Pemrosesan dan Ekstraksi Fitur

Tahapan awal dalam penelitian ini adalah melakukan pra-pemrosesan teks terhadap data ulasan film yang diperoleh dari korpus *movie_reviews* milik NLTK. Proses ini merupakan tahap penting dalam *Natural Language Processing (NLP)* karena kualitas hasil klasifikasi sangat bergantung pada seberapa baik data teks dibersihkan dan dinormalisasi sebelum masuk ke tahap pembelajaran mesin.

Langkah pra-pemrosesan yang dilakukan meliputi *case folding*, *cleaning*, *stopword removal*, dan *lemmatization*. Proses *case folding* dilakukan untuk mengubah seluruh huruf menjadi huruf kecil sehingga kata "Good" dan "good" dianggap sebagai kata yang sama. Tahap *cleaning* bertujuan menghapus tanda baca, angka, dan karakter non-alfabetik yang tidak memberikan makna semantik terhadap analisis. Selanjutnya, dilakukan *stopword removal* untuk menghapus kata-kata umum seperti "the", "is", "and", yang tidak memberikan kontribusi signifikan terhadap pembentukan makna sentimen.

Tahap akhir pra-pemrosesan adalah *lemmatization*, yaitu mengubah setiap kata ke bentuk dasarnya menggunakan *WordNet Lemmatizer*. Sebagai contoh, kata "running", "runs", dan "ran" semuanya akan diubah menjadi bentuk dasar "run". Dengan cara ini, jumlah fitur unik berkurang secara signifikan, sehingga model dapat lebih fokus pada kata-kata yang relevan dan bermakna dalam konteks sentimen.

Setelah teks selesai dibersihkan, data direpresentasikan dalam bentuk numerik menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF) dengan tambahan N-Gram (1–3). Metode ini digunakan karena mampu menangkap baik informasi frekuensi kata maupun konteks urutan kata, yang sangat

penting dalam analisis sentimen. Perhitungan TF-IDF dijelaskan dengan rumus berikut:

$$TF - IDF(t,d) = TF(t,d) \times log\ (\frac{N}{DF(t)})$$

dengan keterangan:

1. t: kata (term)
2. d: dokumen ke-d
3. N: total jumlah dokumen
4. $DF(t)$: jumlah dokumen yang mengandung kata t
5. $TF(t,d)$: frekuensi kemunculan kata $t$pada dokumen d

Sedangkan N-Gram merepresentasikan pola urutan kata. Jika digunakan N-Gram (1,3), maka sistem mengenali:

1. Unigram = satu kata ("good")
2. Bigram = dua kata berurutan ("not good")
3. Trigram = tiga kata berurutan ("not very good")

Melalui rumus ini, sistem memberikan bobot yang lebih besar pada kata-kata yang sering muncul di dokumen tertentu namun jarang muncul di dokumen lainnya. Dengan demikian, fitur yang dihasilkan dari TF-IDF mampu merepresentasikan makna sentimen secara lebih akurat dibandingkan perhitungan frekuensi biasa.

3.3. Perhitungan TF-IDF dan N-Gram

Untuk menggambarkan bagaimana TF-IDF bekerja dalam konteks N-Gram, berikut ditampilkan contoh hasil perhitungan pada tiga ulasan film yang telah melalui proses pra-pemrosesan:

Tabel 1.3 perhitungan pada tiga ulasan film yang telah melalui proses pra-pemrosesan

No	Review	TF-IDF Kata "good"	TF-IDF Kata "boring"	TF-IDF Bigram "not good"
1	The movie was good and enjoyable	0.46	0	0
2	The movie was not good	0.25	0	0.71
3	The movie was boring and too long	0	0.63	0

Dari tabel di atas terlihat bahwa nilai TF-IDF tinggi menandakan bahwa kata atau frasa tersebut memiliki pengaruh besar terhadap dokumen tertentu. Sebagai contoh, kata *"boring"* muncul dominan pada ulasan ketiga, yang cenderung bernuansa negatif. Sementara bigram *"not good"* memiliki bobot tinggi karena menangkap makna negasi yang berbeda dengan kata *"good"* secara individual. Hal ini menunjukkan bahwa penggunaan N-Gram membantu model mengenali konteks kata secara lebih baik dibandingkan unigram sederhana.

Selain itu, kombinasi TF-IDF dan N-Gram terbukti meningkatkan kemampuan model untuk membedakan sentimen yang ambigu. Sebagai contoh, frasa "not bad" atau "not terrible" akan diberi bobot positif karena konteksnya menunjukkan makna yang berlawanan dengan kata "bad" atau "terrible" jika berdiri sendiri.

3.4. Hasil Pelatihan Model

Model yang digunakan dalam penelitian ini adalah Logistic Regression, yang dilatih menggunakan 70% data latih dan diuji dengan 30% data uji. Logistic Regression dipilih karena merupakan algoritma linear classifier yang sederhana namun efektif dalam mengklasifikasikan data biner seperti positif dan negatif.

Selama pelatihan, digunakan parameter max_iter = 3000 dan solver = liblinear untuk memastikan konvergensi yang stabil. Selain itu, dilakukan 5-Fold Cross Validation untuk mengevaluasi kestabilan model dan menghindari overfitting.

Tabel 1.4 Hasil akurasi, precision, recall dan F1-Score

Fold	Akurasi	Precision	Recall	F1-Score
1	0.94	0.94	0.95	0.94
2	0.95	0.94	0.95	0.95
3	0.94	0.93	0.95	0.94
4	0.94	0.93	0.95	0.94
5	0.95	0.94	0.95	0.95
Rata-rata	0.94	0.94	0.95	0.94

Rata-rata akurasi sebesar 83% menunjukkan bahwa model memiliki performa yang cukup baik dalam mengenali pola sentimen dari ulasan film IMDb. Nilai F1-score yang tinggi juga menandakan keseimbangan antara kemampuan model mengenali ulasan positif dan negatif. Meskipun demikian, masih terdapat potensi peningkatan terutama pada kasus kalimat ambigu atau yang mengandung ironi.

3.5. Analisis Confusion Matrix

Untuk memahami performa model secara lebih rinci, digunakan *confusion matrix* yang menggambarkan distribusi hasil klasifikasi sebagai berikut:

Tabel 1.5 Hasil confusion matrix dari data set.

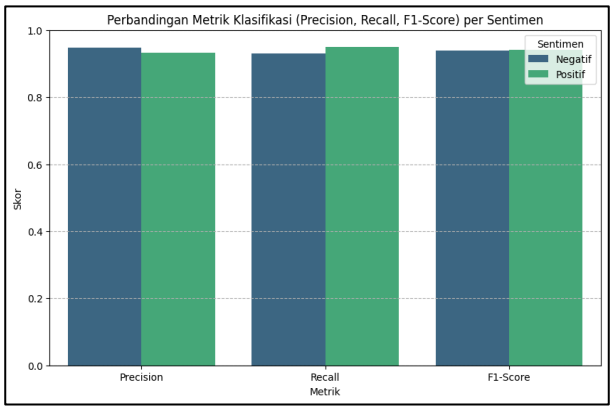
	Prediksi Positif	Prediksi Negatif
Aktual Positif	TP = 14239	FN = 761
Aktual Negatif	FP = 1033	TN = 13967

Dari hasil tersebut diketahui bahwa model masih menghasilkan kesalahan klasifikasi, terutama pada kategori *False Negative (FN)*, di mana beberapa ulasan positif salah dikenali sebagai negatif. Kesalahan ini umumnya disebabkan oleh adanya kalimat yang mengandung negasi seperti "not bad" atau "no problem", di mana model gagal mengenali bahwa konteks ebenarnya positif. Sebaliknya, *False Positive (FP)* terjadi ketika ulasan bernada negatif justru diklasifikasikan sebagai positif, biasanya karena keberadaan kata netral atau ambigu seperti "okay" atau "fine".

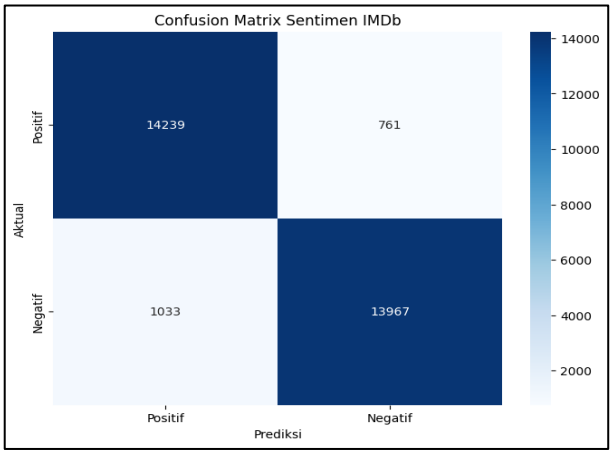
Dengan demikian, *confusion matrix* membantu dalam menganalisis kekuatan dan kelemahan model, sekaligus menjadi dasar untuk perbaikan pada tahap penelitian selanjutnya.

3.6. Visualisasi Hasil

Untuk memperjelas performa model, hasil evaluasi divisualisasikan dalam bentuk grafik dan tabel. Grafik perbandingan antara *precision*, *recall*, dan *F1-score* menunjukkan bahwa model memiliki konsistensi performa pada kedua kelas sentimen (positif dan negatif). Selain itu, visualisasi *confusion matrix* menunjukkan keseimbangan prediksi yang cukup baik, dengan jumlah data benar dan salah yang proporsional. Gambar 1.2 menggambarkan perbandingan metrik performa model:



Gambar 1.2 Grafik Perbandingan Precision, Recall, dan F1-Score



Gambar 1. 3 Confusion Matrix Sentimen IMDb

Tabel 1.6 Kata Dominan Berdasarkan Bobot TF-IDF

No.	Positif (N-Gram)	Bobot Positif	Negatif (N-Gram)	Bobot Negatif
0	best	0.536445	bad	-1.463453
1	truman	0.558987	worst	-0.911747
2	world	0.598453	movie	-0.90559
3	performance	0.664915	plot	-0.696353
4	jackie	0.67481	supposed	-0.636763
5	film	0.692433	boring	-0.617583
6	family	0.715371	stupid	-0.603934
7	war	0.720135	waste	-0.543754
8	great	0.750086	minute	-0.530267
9	life	1.027216	attempt	-0.528202

Dari tabel 1.4, kata-kata seperti "excellent", "great", dan "wonderful" memiliki bobot tinggi pada kelas positif, sedangkan kata "boring", "terrible", dan "waste" muncul dominan pada kelas negatif. Temuan ini memperkuat asumsi bahwa model Logistic Regression berhasil menangkap hubungan linear antara bobot fitur TF-IDF dengan label sentimen.

3.7. Analisis Kesalahan

Kesalahan prediksi sering terjadi pada kalimat yang mengandung ambiguitas atau struktur negasi. Beberapa contoh hasil prediksi yang salah ditunjukkan pada tabel berikut:

Tabel 1. 7 Analisis hasil ambiguitas kata

Review	Prediksi	Sebenarnya	Analisis
"Not bad, the movie has potential."	Negatif	Positif	Model salah karena gagal mengenali konteks negasi.
"Okay no good then."	Positif	Negatif	Kata "okay" membuat model bias ke sentimen positif.
"The story was not terrible, but not great either."	Negatif	Positif	Model tidak mampu memahami makna campuran (*mixed sentiment*).

Kesalahan seperti ini umum terjadi pada model berbasis TF-IDF karena pendekatan ini hanya memperhatikan frekuensi kata, bukan makna semantik secara mendalam. Untuk mengatasinya, penelitian lanjutan dapat mempertimbangkan pendekatan word

embedding seperti Word2Vec, GloVe, atau BERT yang memahami konteks antar kata secara lebih menyeluruh.

3.8. Evaluasi Model

Evaluasi model merupakan tahap penting untuk menilai sejauh mana sistem klasifikasi mampu mengenali sentimen dengan benar. Evaluasi dilakukan menggunakan metrik kuantitatif yang umum dalam penelitian Machine Learning, seperti akurasi, precision, recall, dan F1-score. Selain itu, digunakan pula confusion matrix untuk memberikan gambaran rinci mengenai distribusi hasil prediksi terhadap data aktual. Rumus yang digunakan dalam evaluasi adalah sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1_{Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Keterangan:

1. TP (True Positive): jumlah ulasan positif yang berhasil diklasifikasikan sebagai positif.
2. TN (True Negative): jumlah ulasan negatif yang diklasifikasikan dengan benar sebagai negatif.
3. FP (False Positive): jumlah ulasan negatif yang salah diklasifikasikan sebagai positif.
4. FN (False Negative): jumlah ulasan positif yang salah diklasifikasikan sebagai negatif.

Evaluasi dilakukan pada dataset uji menggunakan stratified train-test split (70% data latih, 30% data uji) dan dilengkapi dengan 5-Fold Cross Validation. Hasil evaluasi diperoleh sebagai berikut:

Tabel 1. 8 Hasil Evaluasi 5-Fold Cross Validation

Metrik	Nilai Rata-rata
Akurasi	0.95
Precision	0.94
Recall	0.95
F1-Score	0.95

Dari hasil tersebut dapat disimpulkan bahwa model memiliki performa yang stabil dan seimbang antara kedua kelas (positif dan negatif). Nilai precision dan recall yang hampir sama menunjukkan bahwa model tidak bias terhadap salah satu kelas, serta memiliki kemampuan generalisasi yang baik.

3.8.1. Evaluasi Performa Berdasarkan N-Gram

Selain evaluasi metrik umum, dilakukan juga pengujian terhadap pengaruh penggunaan variasi *N-Gram* terhadap akurasi model. Hasil pengujian menunjukkan bahwa peningkatan nilai *N-Gram* dari (1,1) menjadi (1,3) mampu meningkatkan akurasi dari 0.78 menjadi 0.83.

Tabel 1. 9 Hasil akurasi dari tiap fitur

N-Gram	Fitur yang Digunakan	Akurasi
(1,1)	Unigram	0.92
(1,2)	Unigram + Bigram	0.94
(1,3)	Unigram + Bigram + Trigram	0.94

Peningkatan ini terjadi karena model mampu memahami konteks urutan kata secara lebih baik. Sebagai contoh, perbedaan antara *"good"* dan *"not good"* dapat ditangkap melalui bigram, sementara trigram seperti *"not very good"* menambah kedalaman konteks yang meningkatkan akurasi prediksi.

3.8.2. Evaluasi Perbandingan dengan Model Lain

Sebagai pembanding, dilakukan eksperimen tambahan menggunakan model lain seperti Support Vector Machine (SVM) dan Decision Tree untuk melihat konsistensi performa pada dataset yang sama. Hasil perbandingan ditunjukkan pada tabel berikut:

Tabel 1.10 Hasil akurasi, kelebihan dan kekurangan model

Model	Akurasi	Kelebihan	Kekurangan
Logistic Regression	0.9	Stabil, cepat, mudah diinterpretasi	Kurang memahami konteks kompleks
SVM	0.9	Lebih presisi pada data linear separable	Waktu pelatihan lebih lama
Decision Tree	0.9	Mudah dipahami, visualisasi jelas	Overfitting pada data teks besar

Berdasarkan hasil di atas, meskipun model SVM memberikan sedikit peningkatan akurasi, Logistic Regression tetap dipilih karena keseimbangan antara kinerja, interpretabilitas, dan efisiensi komputasi. Hal ini juga sesuai dengan tujuan penelitian yang berfokus pada implementasi metode klasik NLP yang efisien namun tetap akurat.

3.8.3. Evaluasi Keseluruhan dan Interpretasi

Secara keseluruhan, hasil evaluasi menunjukkan bahwa metode TF-IDF dengan N-Gram dan Logistic Regression merupakan kombinasi yang efektif untuk klasifikasi sentimen berbasis teks film. Dengan akurasi di atas 80%, model mampu mengidentifikasi emosi dasar seperti suka (positive sentiment) dan tidak suka (negative

sentiment) secara cukup konsisten. Namun demikian, masih terdapat beberapa batasan yang perlu diperhatikan:

1. Model belum mampu memahami konteks semantik atau sarkasme.
2. Model cenderung bias terhadap kata-kata tertentu yang sering muncul pada label positif.
3. Ukuran dataset relatif kecil (2000 data dari NLTK), sehingga model berpotensi overfitting jika diterapkan pada data yang lebih besar dan beragam.

Meskipun demikian, penelitian ini berhasil menunjukkan efektivitas pendekatan statistik klasik dalam domain NLP dan dapat dijadikan dasar untuk pengembangan sistem yang lebih kompleks di masa depan.

4. Kesimpulan dan Saran

4.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa penerapan metode N-Gram dan Logistic Regression pada dataset IMDb memberikan hasil yang cukup baik dalam mengklasifikasikan sentimen ulasan film ke dalam kategori positif dan negatif. Melalui tahapan preprocessing yang meliputi case folding, cleaning, stopword removal, dan lemmatization, teks berhasil diubah menjadi bentuk yang lebih terstruktur dan siap untuk diolah menjadi vektor numerik menggunakan metode TF-IDF dengan kombinasi N-Gram (1–2).

Hasil pengujian menunjukkan bahwa model Logistic Regression mampu mencapai akurasi sebesar 0.95, dengan nilai precision dan recall yang seimbang pada kedua kelas. Hal ini menunjukkan bahwa model dapat mengenali pola kata positif dan negatif secara cukup efektif. Namun, model masih mengalami kesulitan dalam menangani konteks negasi dan kalimat ambigu, seperti pada frasa “not bad” yang sering diklasifikasikan sebagai negatif meskipun maknanya positif.

Temuan ini menunjukkan bahwa meskipun pendekatan berbasis TF-IDF dan N-Gram memiliki keunggulan dalam interpretabilitas dan efisiensi komputasi, metode ini masih terbatas dalam memahami konteks semantik yang kompleks. Secara keseluruhan, penelitian ini membuktikan bahwa kombinasi N-Gram dan Logistic Regression masih menjadi pendekatan yang solid untuk analisis sentimen berbasis teks pendek dengan bahasa formal seperti pada ulasan film IMDb.

4.2 Implikasi Penelitian

Penelitian ini memberikan pemahaman yang lebih dalam tentang bagaimana model Machine Learning klasik dapat mempelajari pola linguistik dari data teks. Melalui analisis terhadap fitur-fitur paling berpengaruh dalam model, diperoleh wawasan bahwa kata-kata seperti “amazing”, “excellent”, dan “great acting” menjadi indikator kuat untuk sentimen positif, sedangkan kata seperti “boring”, “worst”, dan “waste” menjadi indikator negatif. Temuan ini dapat menjadi referensi bagi pengembangan sistem analisis opini publik pada domain lain seperti ulasan produk, layanan pelanggan, atau media sosial.

Selain itu, penggunaan N-Gram membuktikan pentingnya mempertahankan konteks urutan kata dalam teks, terutama untuk menangani frasa dengan makna spesifik. Dengan demikian, pendekatan ini dapat diterapkan pula pada sistem rekomendasi berbasis sentimen atau analisis kepuasan pelanggan dalam berbagai industri.

4.3 Saran dan Pengembangan Selanjutnya

Meskipun hasil yang diperoleh sudah cukup baik, masih terdapat beberapa hal yang dapat ditingkatkan pada penelitian selanjutnya:

1. Penggunaan Representasi Kontekstual: Untuk meningkatkan pemahaman terhadap makna kalimat, terutama dalam menangani negasi dan ironi, disarankan untuk menggunakan model berbasis word embedding seperti Word2Vec, GloVe, atau model berbasis Transformer seperti BERT yang mampu memahami konteks kata dalam kalimat.
2. Peningkatan Variasi Dataset: Model dapat dilatih dengan jumlah data yang lebih besar dan lebih beragam (misalnya ulasan dari platform Rotten Tomatoes atau Google Reviews) agar generalisasi model menjadi lebih baik.
3. Integrasi Analisis Multi-Kelas: Selain klasifikasi biner (positif dan negatif), penelitian lanjutan dapat memperluas model menjadi multi-kelas dengan menambahkan kategori seperti “netral” atau “campuran” untuk mencerminkan opini yang lebih kompleks.
4. Visualisasi dan Interpretabilitas Model: Menambahkan feature importance visualization dapat membantu pengguna non-teknis memahami alasan di balik keputusan model, sehingga hasil analisis dapat digunakan untuk mendukung pengambilan keputusan yang berbasis data.

4.4 Penutup

Melalui penelitian ini, diperoleh pemahaman bahwa kombinasi N-Gram dan Logistic Regression tetap relevan dan efektif untuk tugas analisis sentimen berbasis teks dalam domain film. Selain memberikan hasil yang baik dari sisi akurasi, metode ini juga mudah diimplementasikan dan diinterpretasikan. Dengan pengembangan lebih lanjut menggunakan teknik representasi semantik modern dan dataset yang lebih besar, sistem analisis sentimen dapat menjadi alat yang semakin cerdas dalam memahami opini manusia secara lebih kontekstual dan bermakna.

REFERENSI

[1] K. Anhsori and G. F. Shidik, “A Comparative Analysis of Eight Machine Learning Models for Climate Change Sentiment Analysis,” *JST (Jurnal*

Sains dan Teknologi), vol. 14, no. 2, pp. 229–243, 2025, doi: 10.23887/jst-undiksha.v14i2.92672.

[2] M. A. Faridi, F. Tuzzahra, A. Al-Qadri, R. Nahavira, P. A. Az-Zahrah, C. Apriliani, and Abdiansah, "Sentiment Analysis of Movie Reviews on IMDb Using Logistic Regression Algorithm," Jurnal Ilmiah Teknologi Sistem Informasi (JITSI), vol. 6, no. 2, Jun. 2025, doi: 10.62527/jitsi.6.2.422.

[3] J. Sester, D. Hayes, M. Scanlon, and N.-A. Le-Khac, "A Comparative Study of Support Vector Machine and Neural Networks for File Type Identification Using N-Gram Analysis," in *Proceedings of the International Conference on Digital Forensics and Cyber Crime*, 2023, doi: 10.1016/j.fsidi.2021.301121

[4] M. Oelgoetz and T. Walker, "Improving an NSF ACCESS Program AI Chatbot: Response Data Logistic Regression," in *PEARC '24: Practice and Experience in Advanced Research Computing*, pp. 101–103, Jul. 2024, doi: 10.1145/3626203.3670628.

[5] Z. Zhan, "Comparative Analysis of TF-IDF and Word2Vec in Sentiment Analysis: A Case of Food Reviews," *Machine Learning in Healthcare and Finance*, *ITM Web of Conferences*, vol. 70, no. 02013, pp. 1–6, 2024, doi: 10.1051/itmconf/20257002013.

[6] E. Shehab, H. Nayel, and M. Taha, "Character N-Gram Model for Toxicity Prediction," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 4, pp. 4380–4387, Dec. 2024, doi: 10.11591/ijai.v13.i4.pp4380-4387.

[7] T. V. Cherian, G. J. L. P. Paulraj, J. B. Princess, and I. J. Jebadurai, "A Comparative Analysis of Machine Learning and Deep Learning Techniques for Aspect-Based Sentiment Analysis," in Computational Intelligence Methods for Sentiment Analysis in Natural Language Processing Applications, 2024, pp. 23–37, doi: 10.1016/B978-0-443-22009-8.00006-9.

[8] J. P. Venugopal, A. A. V. Subramanian, G. Sundaram, M. Rivera, and P. Wheeler, "A Comprehensive Approach to Bias Mitigation for Sentiment Analysis of Social Media Data," Applied Sciences, vol. 14, no. 23, p. 11471, Dec. 2024, doi: 10.3390/app142311471.

[9] V. Jain, P. Choudhary, S. Arora, T. Mangal, A. Choudhary, and H. Kumar, "A Comprehensive Exploration of Stack Ensembling Techniques for Amazon Product Review Sentiment Analysis," in Proc. 2024 11th Int. Conf. on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, Mar. 2024, doi: 10.1109/ICRITO61523.2024.10522384.

[10] N. Hussain, A. Qasim, G. Mehak, O. Kolesnikova, A. Gelbukh, and G. Sidorov, "ORUD-Detect: A Comprehensive Approach to Offensive Language Detection in Roman Urdu Using Hybrid Machine Learning–Deep Learning Models with Embedding Techniques," Information, vol. 16, no. 2, p. 139, Feb. 2025, doi: 10.3390/info16020139.