Nama : Rudi Hartadi Setiawan

Nim : A11.2022.14081

Link Drive : pba

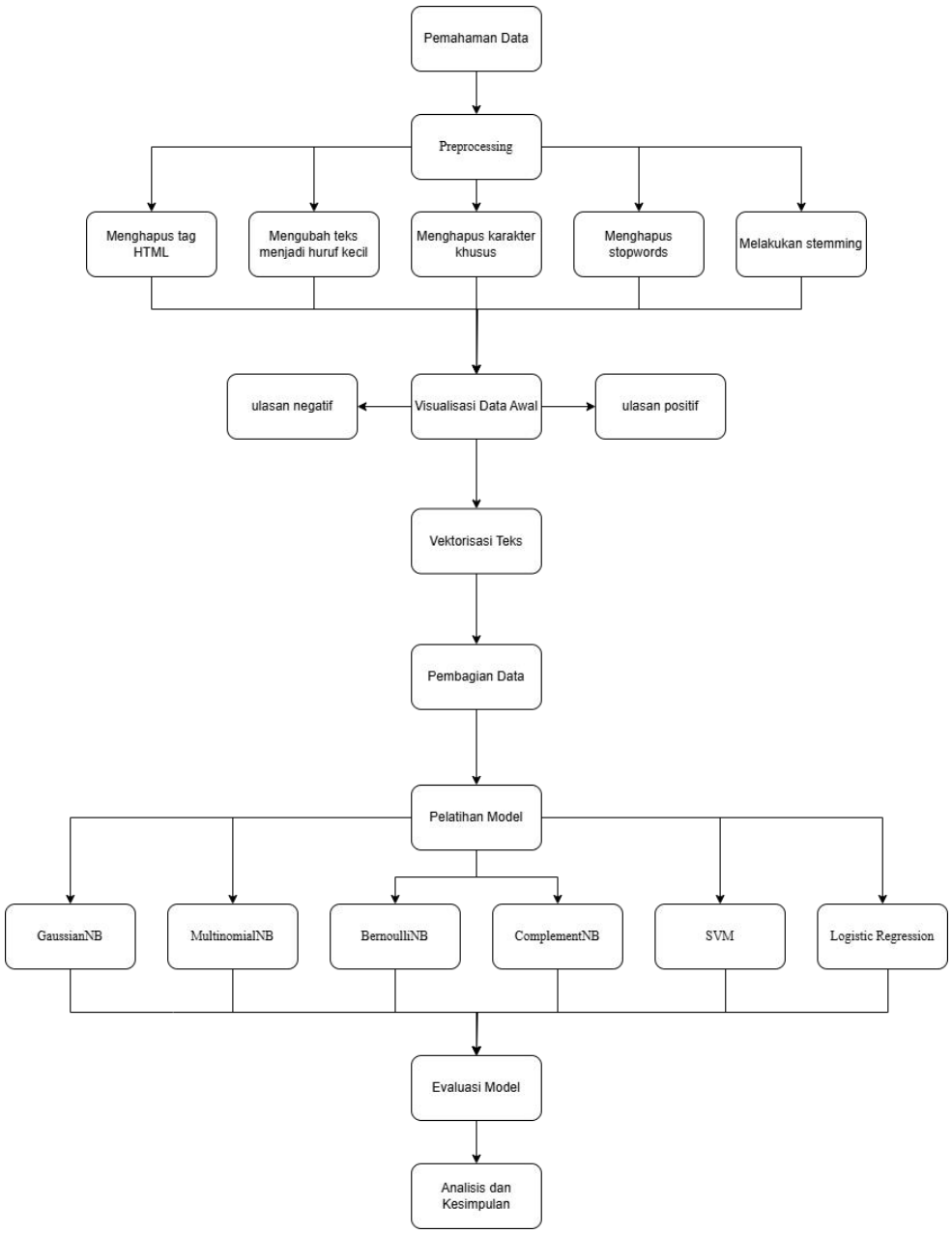
Link Video : Salinan bym-axkf-vxx (2025-07-03 12:16 GMT+7)

Link Dataset : dataset

Link Source Code: nlp.ipynb

Laporan Analisis
Sentimen Ulasan Film IMDB

Analisis sentimen, atau opinion mining, adalah bidang dalam Natural Language Processing (NLP) yang bertujuan untuk mengidentifikasi, mengekstrak, dan mempelajari sentimen yang mendasari suatu teks. Dalam laporan ini, kami menganalisis dataset ulasan film dari IMDB untuk mengklasifikasikan sentimen ulasan sebagai positif atau negatif. Tujuan utama proyek ini adalah untuk membandingkan kinerja berbagai model machine learning pada data tekstual setelah serangkaian tahap preprocesing yang teliti.



1. Deskripsi Dataset

Dataset yang menjadi fokus analisis ini adalah "IMDB Dataset.csv". Ini merupakan kumpulan ulasan film yang sering digunakan dalam tugas-tugas analisis sentimen.

A. Ukuran Awal:

```
[5] df.shape

(50000, 2)
```

Dataset lengkap mencakup 50.000 ulasan film.

B. Pengambilan Sampel:

```
<class 'pandas.core.frame.DataFrame'>
Index: 10000 entries, 6854 to 29185
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   review     10000 non-null  object
 1   sentiment  10000 non-null  object
dtypes: object(2)
memory usage: 234.4+ KB
```

Untuk menjaga efisiensi komputasi dan fokus analisis, sebuah subset acak sebanyak 10.000 ulasan dipilih dari dataset asli. Hal ini memungkinkan eksplorasi model tanpa beban komputasi yang berlebihan, sekaligus tetap mempertahankan representasi yang cukup dari distribusi sentimen keseluruhan.

C. Struktur Data:

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive

Dataset ini memiliki dua kolom utama:

a) review: Berisi teks ulasan film yang bervariasi dalam panjang dan kompleksitasnya. Contoh ulasan dapat dilihat pada notebook yang disediakan.

b) sentiment: Label kelas yang menunjukkan polaritas sentimen ulasan, yaitu "positive" atau "negative".

D. Distribusi Sentimen:

sentiment	count
positive	5055
negative	4945

dtype: int64

Setelah pengambilan sampel, distribusi sentimen dalam dataset adalah sebagai berikut:

a) Positif: 5055 ulasan (50.55%)

b) Negatif: 4945 ulasan (49.45%)

Distribusi yang hampir seimbang ini menunjukkan bahwa dataset tidak mengalami imbalance, yang merupakan karakteristik ideal untuk melatih model klasifikasi biner. Sentimen "positive" dikodekan sebagai 1 dan "negative" sebagai 0 untuk keperluan pemodelan numerik.

2. Preprocessing Data Tekstual

Preprocessing teks adalah tahap krusial dalam NLP yang mengubah teks mentah menjadi representasi yang lebih bersih dan terstruktur untuk analisis. Langkah-langkah yang diterapkan di sini dirancang untuk mengurangi noise dan meningkatkan kualitas fitur yang diekstraksi.

A. Penghapusan Tag HTML: Ulasan film sering kali mengandung tag HTML seperti
 yang tidak memberikan informasi sentimen dan dapat mengganggu analisis. Oleh karena itu, semua tag HTML dihapus dari setiap ulasan menggunakan ekspresi reguler.

B. Konversi ke Huruf Kecil (Lowercasing): Semua teks diubah menjadi huruf kecil. Ini menghindari perlakuan kata yang sama (misalnya, "Film" dan "film") sebagai entitas yang berbeda, sehingga mengurangi kompleksitas kosakata dan meningkatkan efisiensi vektorisasi.

C. Penghapusan Karakter Khusus: Karakter non-alfanumerik (seperti @, %, *, dll.) diubah menjadi spasi kosong. Ini membantu memfokuskan analisis pada kata-kata inti dan menghindari noise dari simbol yang tidak relevan.

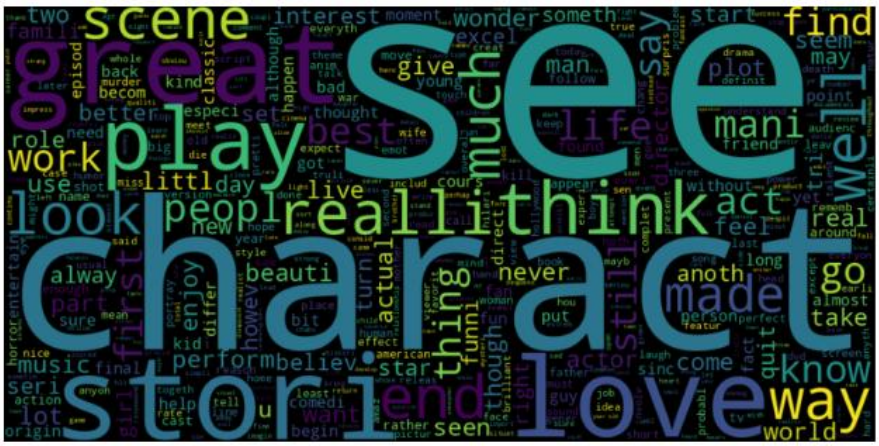
D. Penghapusan Stopwords: Stopwords adalah kata-kata yang sangat umum dalam suatu bahasa (misalnya, "the", "a", "is") yang biasanya tidak membawa makna sentimen yang signifikan. Menggunakan daftar stopwords dari NLTK, kata-kata ini dihapus untuk meningkatkan rasio signal-to-noise dalam data. Kata-kata seperti 'movi', 'film', 'word', 'time', 'one', 'make', 'watch', 'show', 'good', dan 'even' juga secara spesifik dihapus karena dianggap kurang memberikan kontribusi sentimen yang signifikan dalam konteks ulasan film.

E. Stemming (Porter Stemmer): Stemming adalah teknik untuk mengurangi kata menjadi bentuk akarnya atau kata dasarnya. Misalnya, "loved", "loving", dan "loves" semuanya akan direduksi menjadi "love". PorterStemmer diterapkan untuk mengurangi infleksi kata, yang membantu mengurangi jumlah fitur unik dan meningkatkan generalisasi model.

3. Visualisasi Data

Untuk mendapatkan wawasan awal tentang karakteristik linguistik dari ulasan positif dan negatif, kami menggunakan word clouds.

A. Word Cloud Sentimen Positif:



Word cloud untuk ulasan positif menampilkan kata-kata yang paling sering muncul dalam ulasan yang dinilai positif. Kata-kata seperti "great", "love", "best", "perfect", "enjoy", dan "amaz" (dari stemming "amazing") kemungkinan akan

menonjol, memberikan gambaran sekilas tentang kosakata yang terkait dengan sentimen baik.

B. Word Cloud Sentimen Negatif:

Sebaliknya, word cloud untuk ulasan negatif akan menyoroti kata-kata yang lebih sering digunakan dalam ulasan dengan sentimen negatif. Kata-kata seperti "bad", "worst", "disappoint", "terribl", dan "wast" (dari stemming "waste") diharapkan muncul lebih besar, mengindikasikan dominasi istilah-istilah ini dalam konteks ulasan negatif.

Word clouds yang dihasilkan dari notebook mengonfirmasi adanya perbedaan yang jelas dalam kosakata yang digunakan untuk setiap polaritas sentimen, meskipun beberapa kata netral mungkin muncul di kedua kategori.

4. Metode dan Eksperimen

4.1. Vektorisasi Teks

Setelah preprocesing, teks perlu diubah menjadi format numerik agar dapat digunakan oleh model machine learning. Bag-of-Words (BoW) adalah metode yang digunakan di sini.

a) CountVectorizer: CountVectorizer dari Scikit-learn digunakan untuk mengonversi kumpulan dokumen teks menjadi matriks token counts. Setiap baris dalam matriks ini mewakili sebuah ulasan, dan setiap kolom mewakili sebuah kata unik dalam korpus.

b) Pembatasan Fitur: Parameter max_features=2500 diterapkan pada CountVectorizer. Ini berarti hanya 2500 kata yang paling sering muncul

(setelah preprocesing) yang akan digunakan sebagai fitur. Pembatasan ini membantu mengurangi dimensi data, mencegah overfitting, dan meningkatkan efisiensi komputasi. Data hasil vektorisasi memiliki bentuk (10000, 2500).

4.2. Pembagian Data

Dataset yang telah divetorisasi (x) dan label sentimen (y) kemudian dibagi menjadi set pelatihan (training set) dan set pengujian (testing set).

A. train_test_split: Fungsi ini digunakan untuk membagi data.

B. Ukuran Pengujian: test_size=0.2 menunjukkan bahwa 20% dari data akan dialokasikan untuk pengujian, dan 80% sisanya untuk pelatihan. Ini menghasilkan:

   a) x_train: 8000 sampel untuk pelatihan fitur.

   b) y_train: 8000 sampel untuk pelatihan label.

   c) x_test: 2000 sampel untuk pengujian fitur.

   d) y_test: 2000 sampel untuk pengujian label.

4.3. Model yang Digunakan

Beberapa model klasifikasi machine learning dilatih dan dievaluasi:

A. Gaussian Naive Bayes (GaussianNB): Sebuah model Naive Bayes yang mengasumsikan distribusi fitur Gaussian. Model ini sering digunakan sebagai baseline karena kesederhanaannya.

B. Multinomial Naive Bayes (MultinomialNB): Ideal untuk klasifikasi dokumen dengan fitur yang mewakili frekuensi hitungan (seperti Bag-of-Words).

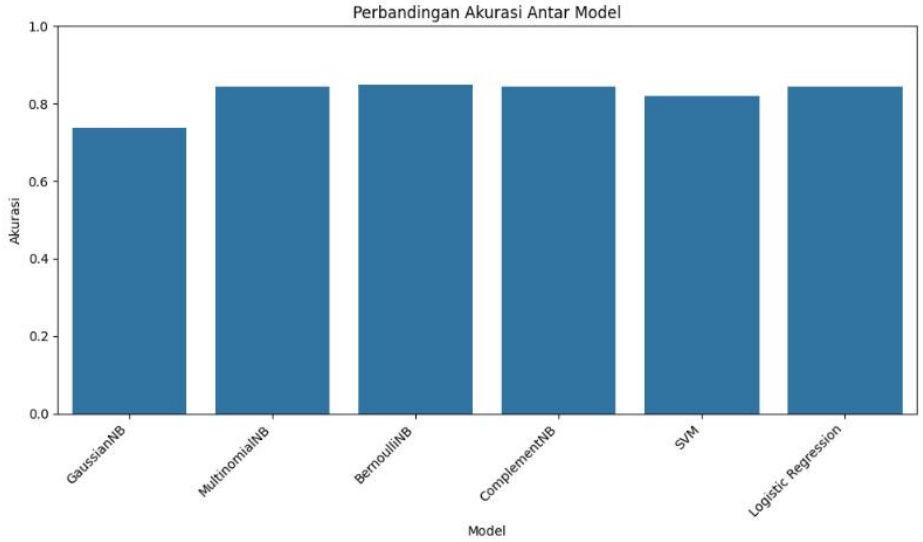
C. Bernoulli Naive Bayes (BernoulliNB): Cocok untuk fitur biner (ada atau tidak ada suatu kata).

D. Complement Naive Bayes (ComplementNB): Varian Naive Bayes yang dirancang untuk menangani dataset dengan distribusi kelas yang tidak seimbang.

E. Support Vector Machine (SVM): Model yang kuat yang mencari hyperplane optimal untuk memisahkan kelas. Digunakan dengan kernel 'linear'.

F. Regresi Logistik (Logistic Regression): Model linier yang digunakan untuk masalah klasifikasi biner, menghitung probabilitas suatu instans termasuk dalam kelas tertentu.

5. Hasil Eksperimen dan Evaluasi



Setiap model dievaluasi berdasarkan metrik kinerja standar dalam klasifikasi: Akurasi, Presisi, Recall, dan F1-score. Matriks kebingungan (confusion matrix) juga disediakan untuk setiap model.

Model	Akurasi	Presisi	Recall	F1-score
GaussianNB	0.7390	0.835196	0.596806	0.696158
MultinomialNB	0.8430	0.861345	0.818363	0.839304
BernoulliNB	0.8485	0.856997	0.837325	0.847047
ComplementNB	0.8430	0.862105	0.817365	0.839139
SVM	0.8205	0.825076	0.814371	0.819689
Logistic Regression	0.8430	0.843313	0.843313	0.843313

5.1. Analisis Hasil:

A. BernoulliNB:

```
Bernoulli Naive Bayes Model:
Accuracy: 0.8485
Confusion Matrix:
 [[858 140]
 [163 839]]
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.86      0.85       998
           1       0.86      0.84      0.85      1002

    accuracy                           0.85      2000
   macro avg       0.85      0.85      0.85      2000
weighted avg       0.85      0.85      0.85      2000
```

Model BernoulliNB menunjukkan akurasi tertinggi sebesar 84.85%. Ini sangat relevan karena BernoulliNB cocok untuk fitur biner, di mana kehadiran atau ketiadaan kata lebih diutamakan daripada frekuensi absolutnya dalam model Bag-of-Words.

B. Kinerja Konsisten:

a) MultinomialNB

```
Multinomial Naive Bayes Model:
Accuracy: 0.843
Confusion Matrix:
 [[866 132]
 [182 820]]
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.87      0.85       998
           1       0.86      0.82      0.84      1002

    accuracy                           0.84      2000
   macro avg       0.84      0.84      0.84      2000
weighted avg       0.84      0.84      0.84      2000
```

b) ComplementNB

```
Complement Naive Bayes Model:
Accuracy: 0.843
Confusion Matrix:
 [[867 131]
 [183 819]]
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.87      0.85       998
           1       0.86      0.82      0.84      1002

    accuracy                           0.84      2000
   macro avg       0.84      0.84      0.84      2000
weighted avg       0.84      0.84      0.84      2000
```

c) Logistic Regression

```
Logistic Regression Model:
Accuracy: 0.843
Confusion Matrix:
 [[841 157]
 [157 845]]
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.84      0.84       998
           1       0.84      0.84      0.84      1002

    accuracy                           0.84      2000
   macro avg       0.84      0.84      0.84      2000
weighted avg       0.84      0.84      0.84      2000
```

Model-model tersebut menunjukkan kinerja yang sangat baik dan konsisten, dengan akurasi sekitar 84.30%. Hal ini menunjukkan bahwa model-model ini, yang memang sering digunakan dalam klasifikasi teks berbasis frekuensi, bekerja dengan efektif pada dataset ini.

C. GaussianNB:

```
Gaussian Naive Bayes Model:
Accuracy: 0.739
Confusion Matrix:
 [[880 118]
 [404 598]]
Classification Report:
              precision    recall  f1-score   support

           0       0.69      0.88      0.77       998
           1       0.84      0.60      0.70      1002

    accuracy                           0.74      2000
   macro avg       0.76      0.74      0.73      2000
weighted avg       0.76      0.74      0.73      2000
```

Model ini memiliki akurasi terendah (73.90%). Ini sesuai dengan ekspektasi karena GaussianNB mengasumsikan distribusi fitur yang kontinu dan berbentuk Gaussian, yang kurang sesuai untuk fitur diskrit dan jarang seperti hitungan kata dalam model Bag-of-Words. Model ini memiliki recall yang tinggi untuk kelas 0 (ulasan negatif) tetapi recall yang lebih rendah untuk kelas 1 (ulasan positif), menunjukkan kecenderungan untuk lebih sering memprediksi negatif.

D. SVM:

```
SVM Model:
Accuracy: 0.8205
Confusion Matrix:
 [[825 173]
 [186 816]]
Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.83      0.82       998
           1       0.83      0.81      0.82      1002

    accuracy                           0.82      2000
   macro avg       0.82      0.82      0.82      2000
weighted avg       0.82      0.82      0.82      2000
```

Meskipun dikenal sebagai model yang kuat, SVM dengan kernel linear menunjukkan akurasi sedikit lebih rendah (82.05%) dibandingkan model Naive Bayes dan Regresi Logistik teratas. Ini mungkin karena kompleksitas komputasi yang lebih tinggi untuk menemukan hyperplane optimal pada dimensi fitur yang cukup besar, atau karena sifat linier kernel yang mungkin tidak sepenuhnya menangkap hubungan non-linier dalam data.

5.2. Diskusi Performa Model

Dalam proyek analisis sentimen ini, model Bernoulli Naive Bayes (BernoulliNB) menunjukkan performa terbaik dengan akurasi 84.85%. Model ini diikuti oleh Multinomial Naive Bayes (MultinomialNB), Complement Naive Bayes (ComplementNB), dan Logistic Regression, yang semuanya memiliki akurasi sekitar 84.30%. Model Gaussian Naive Bayes (GaussianNB) menunjukkan akurasi terendah, sementara Support Vector Machine (SVM) berada di tengah.

A. Performa unggul BernoulliNB pada dataset ini dapat dijelaskan oleh sifat fitur Bag-of-Words (BoW) yang digunakan dan asumsi model itu sendiri:

1) Representasi Biner yang Sesuai: BernoulliNB secara inheren dirancang untuk bekerja dengan fitur biner. Dalam model Bag-of-Words, meskipun kita menghitung frekuensi kata, CountVectorizer pada dasarnya menciptakan representasi di mana kehadiran atau ketiadaan sebuah kata (token) dalam dokumen adalah fitur yang paling dominan. BernoulliNB memodelkan probabilitas kehadiran suatu fitur biner (1 jika kata ada, 0 jika tidak ada) untuk setiap kelas. Untuk klasifikasi sentimen, seringkali yang lebih penting adalah apakah kata-kata kunci sentimen (misalnya, "hebat", "buruk") hadir atau tidak, daripada seberapa sering kata-kata tersebut berulang dalam satu ulasan.

2) Mengatasi Data Sparse: Data Bag-of-Words seringkali sangat sparse (banyak nilai nol), karena sebagian besar ulasan hanya mengandung sebagian kecil dari total kosakata. BernoulliNB menangani sparsity ini dengan baik karena hanya mempertimbangkan keberadaan fitur, bukan frekuensinya.

3) Kesederhanaan dan Efisiensi: Seperti model Naive Bayes lainnya, BernoulliNB relatif sederhana dan efisien secara komputasi, menjadikannya pilihan yang cepat untuk dataset tekstual besar.

B. Performa Model Lainnya:

1) Multinomial Naive Bayes (MultinomialNB) & Complement Naive Bayes (ComplementNB): Kedua model ini juga bekerja sangat baik. MultinomialNB secara tradisional bagus untuk data hitungan, memodelkan probabilitas frekuensi kata. ComplementNB, sebagai varian yang mengatasi class imbalance (meskipun dataset ini cukup seimbang), juga menunjukkan kinerja yang setara. Keduanya efektif karena sifatnya yang probabilistik dan asumsi independensi fitur yang sederhana, yang berfungsi dengan baik pada data teks setelah preprocesing.

2) Logistic Regression: Model ini adalah model linier yang kuat untuk klasifikasi biner dan sering menjadi baseline yang tangguh. Akurasi 84.30% menunjukkan kemampuannya dalam memisahkan kelas sentimen secara linier berdasarkan fitur Bag-of-Words.

3) Support Vector Machine (SVM): SVM dengan kernel linear menunjukkan akurasi 82.05%. Meskipun SVM adalah model yang sangat kuat, pada dataset dengan dimensi fitur tinggi seperti Bag-of-Words, kernel linear mungkin terbatas dalam menangkap hubungan kompleks antar fitur. Kernel non-linear (seperti RBF) bisa jadi lebih baik, tetapi akan meningkatkan kompleksitas dan waktu komputasi secara signifikan.

4) Gaussian Naive Bayes (GaussianNB): Model ini memiliki performa terendah (73.90%). GaussianNB mengasumsikan bahwa fitur-fitur mengikuti distribusi Gaussian (normal), yang cocok untuk data kontinu. Namun, fitur-fitur yang berasal dari CountVectorizer (frekuensi kata) adalah diskrit dan seringkali tidak mengikuti distribusi Gaussian. Ketidaksesuaian asumsi ini dengan sifat data menyebabkan performa yang kurang optimal.

C. Implikasi untuk Analisis Sentimen:

Dalam tugas analisis sentimen berbasis Bag-of-Words, model Naive Bayes, khususnya BernoulliNB, seringkali menjadi pilihan yang sangat efektif dan efisien. Ini karena fokusnya pada keberadaan kata-kata kunci sentimen daripada frekuensinya, serta kemampuannya menangani data sparse. Regresi Logistik juga merupakan alternatif yang solid karena interpretasinya yang lebih langsung dan kemampuannya untuk mengidentifikasi fitur yang paling berpengaruh.

6. Kesimpulan

Proyek analisis sentimen pada dataset IMDB ini berhasil menunjukkan efektivitas berbagai model machine learning dalam mengklasifikasikan sentimen ulasan film.

Tahap preprocesing data tekstual, yang meliputi penghapusan tag HTML, lowercasing, penghapusan karakter khusus, penghapusan stopwords, dan stemming, terbukti sangat penting dalam menyiapkan data untuk analisis. Visualisasi word clouds juga memberikan wawasan kualitatif yang berharga mengenai perbedaan kosakata antara sentimen positif dan negatif.

Dari eksperimen model, BernoulliNB muncul sebagai model dengan kinerja terbaik dalam hal akurasi pada dataset yang telah divetorisasi menggunakan Bag-of-Words.

Model-model lain seperti MultinomialNB, ComplementNB, dan Logistic Regression juga menunjukkan hasil yang sangat kompetitif. Ini menegaskan bahwa model-model Naive Bayes, dengan asumsi independensi fitur yang kuat, sering kali sangat efektif dan efisien untuk tugas klasifikasi teks.