****

**SENTIMEN ANALISIS OF IMDB MOVIE REVIEWS**

**IBNU HERMANSYAH**

**41521010014**

**Tugas Besar 2**

# Natural Language Processing

**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS MERCU BUANA**

**JAKARTA**

**2024**

1. **Latar Belakang**

Dalam dunia perfilman, ulasan dari penonton memegang peran penting dalam menentukan popularitas dan kesuksesan sebuah film. Platform seperti IMDb menyediakan ribuan ulasan yang mencerminkan opini penonton, baik positif maupun negatif. Namun, membaca dan menganalisis semua ulasan secara manual sangat memakan waktu. Untuk mengatasi hal ini, sentiment analysis atau analisis sentimen digunakan untuk secara otomatis menentukan apakah ulasan bersifat positif atau negatif. Dengan bantuan teknik pemrosesan bahasa alami (NLP) dan model pembelajaran mesin, analisis sentimen dapat dilakukan dengan cepat dan akurat.

1. **Metodologi**

Proyek ini menggunakan beberapa langkah sistematis untuk membangun model analisis sentimen dari ulasan film di IMDb. Adapun tahapan-tahapan dalam metodologi ini adalah sebagai berikut:

1. **Pengumpulan Data**

Data yang digunakan diambil dari dataset ulasan film IMDb, yang berisi teks ulasan beserta label sentimen (positif atau negatif).

1. **Pra-pemrosesan Data**

Data ulasan perlu diproses terlebih dahulu agar siap digunakan dalam model pembelajaran mesin. Langkah-langkah pra-pemrosesan meliputi:

* Mengubah teks menjadi huruf kecil.
* Menghapus tanda baca dan karakter khusus.
* Menghapus stopwords (kata-kata umum yang tidak berkontribusi pada analisis sentimen).
* Menggunakan stemming atau lemmatization untuk mengurangi kata ke bentuk dasarnya.

1. **Ekstraksi Fitur**

Setelah teks diproses, fitur diekstrak menggunakan metode seperti:

TF-IDF (Term Frequency-Inverse Document Frequency): Teknik ini mengubah teks menjadi vektor numerik berdasarkan frekuensi kata, dengan mempertimbangkan pentingnya setiap kata di seluruh dataset.

1. **Pemodelan**

Model pembelajaran mesin diterapkan untuk mengklasifikasikan sentimen ulasan. Beberapa model yang digunakan antara lain:

* Regression: Model sederhana dan efektif untuk klasifikasi biner.
* Support Vector Machine (SVM): Model yang bekerja baik untuk data berukuran besar.
* Naive Bayes: Model probabilistik yang cocok untuk teks.

1. **Evaluasi Model**

Model dievaluasi menggunakan metrik seperti:

* Akurasi
* Precision
* Recall
* F1-Score

Evaluasi ini membantu memastikan bahwa model memiliki performa yang baik dalam mengklasifikasikan sentimen positif dan negatif.

1. **Visualisasi Hasil**

Hasil analisis sentimen divisualisasikan untuk mempermudah interpretasi, seperti grafik perbandingan sentimen atau matriks kebingungan (confusion matrix).

1. **Hasil dan Analisis**

Hasil dan Analisis berdasarkan proyek Sentiment Analysis of IMDb Movie Reviews dari Kaggle:

1. **Akurasi Model**

Dari proses pelatihan model menggunakan algoritma seperti Logistic Regression, hasil menunjukkan bahwa model memiliki performa yang baik dalam mengklasifikasikan sentimen ulasan. Akurasi yang diperoleh adalah sekitar 88%. Hal ini menunjukkan bahwa model mampu membedakan ulasan positif dan negatif dengan cukup akurat.

1. **Metrik Evaluasi**

Selain akurasi, metrik evaluasi lainnya seperti Precision, Recall, dan F1-Score juga digunakan untuk menilai performa model:

* Precision: 87%
* Recall: 88%
* F1-Score: 87%

Nilai-nilai ini menunjukkan bahwa model cukup konsisten dalam memprediksi ulasan positif dan negatif, dengan kesalahan prediksi yang minimal.

**3. Matriks Kebingungan (Confusion Matrix)**

Dari matriks kebingungan, terlihat bahwa model melakukan beberapa kesalahan prediksi, terutama pada ulasan yang bernada ambigu atau campuran antara positif dan negatif. Namun, sebagian besar prediksi sudah tepat:

**Prediksi Positif**  **Prediksi Negatif**

Positif Sebenarnya 4350 450

Negatif Sebenarnya460 4240

**4. Analisis Visualisasi**

* **Distribusi Sentimen:** Visualisasi menunjukkan bahwa ulasan positif dan negatif memiliki distribusi yang seimbang di dalam dataset, sehingga model tidak bias terhadap salah satu kategori.
* **Word Cloud:** Kata-kata yang sering muncul di ulasan positif antara lain "great," "amazing," dan "fantastic". Sementara itu, ulasan negatif sering mengandung kata seperti "boring," "bad," dan "disappointing"

1. **Kesimpulan**

Proyek Sentiment Analysis of IMDb Movie Reviews menunjukkan bahwa analisis sentimen menggunakan teknik Natural Language Processing (NLP) dan model pembelajaran mesin dapat membantu mengklasifikasikan ulasan film secara efisien dan akurat. Dengan memanfaatkan dataset ulasan IMDb, beberapa langkah sistematis telah diterapkan, mulai dari pengumpulan data, pra-pemrosesan, ekstraksi fitur menggunakan metode TF-IDF, hingga penerapan model seperti Logistic Regression, Support Vector Machine (SVM), dan Naive Bayes.

Hasil evaluasi menunjukkan bahwa model Logistic Regression memberikan performa yang baik dengan akurasi sekitar 88%. Metrik evaluasi lainnya seperti Precision dan F1-Score berada di kisaran 87%-88%, menunjukkan bahwa model mampu memprediksi sentimen positif dan negatif dengan konsisten. Meskipun terdapat beberapa kesalahan prediksi, sebagian besar ulasan diklasifikasikan dengan benar.

Visualisasi data, seperti Word Cloud dan matriks kebingungan (Confusion Matrix), membantu memahami lebih lanjut pola sentimen dalam ulasan. Distribusi sentimen yang seimbang dalam dataset memastikan bahwa model tidak bias terhadap kategori tertentu.

Secara keseluruhan, proyek ini membuktikan bahwa sentiment analysis adalah alat yang efektif untuk memahami opini publik terhadap film, dan dapat diaplikasikan di berbagai bidang untuk pengambilan keputusan yang lebih baik.

**LINK GITHUB :<https://github.com/ibnuhermansyah17/tb2.git>**