Perbandingan Logistic Regression dengan Random Forest dalam Memprediksi Sentimen Pada IMDb Moview Review

Nandi Agung Permana[#1], Hendra Bunyamin[*2]

[#] *Teknik Informatika ,Universitas Kristen Maranatha*

Jl. Surya Sumantri No.65, Sukawarna, Kec. Sukajadi, Kota Bandung

[1]nandiagung8@gmail.com

[2] hendra.bunyamin@it.maranatha.edu

Abstract — **This study aims to compare the performance of Logistic Regression and Random Forest models in predicting sentiment on IMDb Movie Review. Using a dataset consisting of movie reviews, both models are trained and evaluated using commonly used performance evaluation metrics. The research findings indicate that Logistic Regression outperforms Random Forest in predicting sentiment in IMDb reviews. These findings provide valuable insights for practitioners and researchers in selecting the most suitable model for sentiment analysis tasks on movie review data.**

Keywords— **movie review, machine learning, logistic regression, random forest, TF-IDF**

I. PENDAHULUAN

Teknologi internet yang terus berkembang telah menjadi bagian penting dalam kehidupan banyak orang, memungkinkan komunikasi dan pertukaran informasi melalui media sosial. Salah satu bentuk ekspresi di internet adalah ulasan film. Film, sebagai seni yang menggabungkan video, suara, dan cerita, sering kali digunakan untuk menyampaikan pesan dan menjadi sarana hiburan yang umum. Fithratullah (2019) menyatakan bahwa film adalah karya seni yang dibuat berdasarkan kebutuhan dan keinginan masyarakat, dan popularitasnya sering ditentukan oleh ulasan film.Kemajuan teknologi memungkinkan masyarakat berbagi pendapat tentang film di situs jejaring sosial, menjadikan media sosial sumber instan untuk mendapatkan opini publik. Dalam analisis sentimen film, pendapat penonton dapat digunakan untuk menentukan respons mereka terhadap film tersebut, baik positif maupun negatif. Teixeira et al. (2020) menekankan bahwa film adalah proses menciptakan karakter dari peristiwa yang dapat terjadi dalam ruang dan waktu tertentu, sehingga preferensi individu bisa berbeda-beda.Analisis sentimen, sebagai bagian dari pemrosesan bahasa alami (NLP) dan data mining, digunakan untuk mengklasifikasikan film berdasarkan ulasan. Menurut Purnomoputra et al. (2019), metode ini melibatkan komputasi untuk mengidentifikasi perasaan atau ekspresi audiens. Dang et al. (2020) menambahkan bahwa data untuk analisis sentimen biasanya berasal dari media sosial. Hasil analisis sentimen dapat membantu pengambilan keputusan tentang kualitas layanan.IMDb, salah satu platform terkemuka untuk informasi tentang film, menyediakan ulasan, peringkat, dan komentar pengguna yang bermanfaat untuk analisis sentimen. Ulasan dan peringkat yang terorganisir dari IMDb memberikan gambaran langsung tentang reaksi penonton terhadap film. Data ini sangat penting untuk mengevaluasi popularitas, kualitas, dan penerimaan film oleh penonton.Oleh karena itu, ulasan IMDb menjadi bagian penting dari penelitian dan analisis di bidang analisis sentimen film. Internet dan media sosial memungkinkan orang berbagi pendapat mereka tentang film, dan analisis sentimen dari teks ulasan ini sangat penting bagi industri film untuk menilai kualitas film berdasarkan tanggapan penonton. Penelitian ini juga dapat mencakup pengembangan teknik analisis sentimen yang lebih canggih dan penerapan temuan ini dalam praktik industri film.

Rumusan masalah tersebut mencakup evaluasi kinerja kedua model, analisis perbedaan antara keduanya, serta pertimbangan terkait faktor-faktor yang dapat memengaruhi hasil prediksi sentimen pada ulasan film di IMDB.

1. Bagaimana melakukan prediksi sentimen pada IMDB Movie Review
2. Bagaimana kinerja model Random Forest dan Logistic Regression dalam memprediksi sentimen pada IMDb Movie Review?
3. Apakah terdapat perbedaan signifikan antara kinerja Logistic Regression dan Random Forest dalam memprediksi sentimen pada IMDb Movie Review?

4. Bagaimana cara menggunakan model Logistic Regression dan Random Forest dalam antarmuka web?

Penelitian ini bertujuan untuk membandingkan kinerja model Logistic Regression dan Random Forest dalam memprediksi sentimen pada ulasan film IMDb, serta mengimplementasikan web sebagai antarmuka untuk kedua model tersebut. Adapun tujuan spesifiknya adalah:

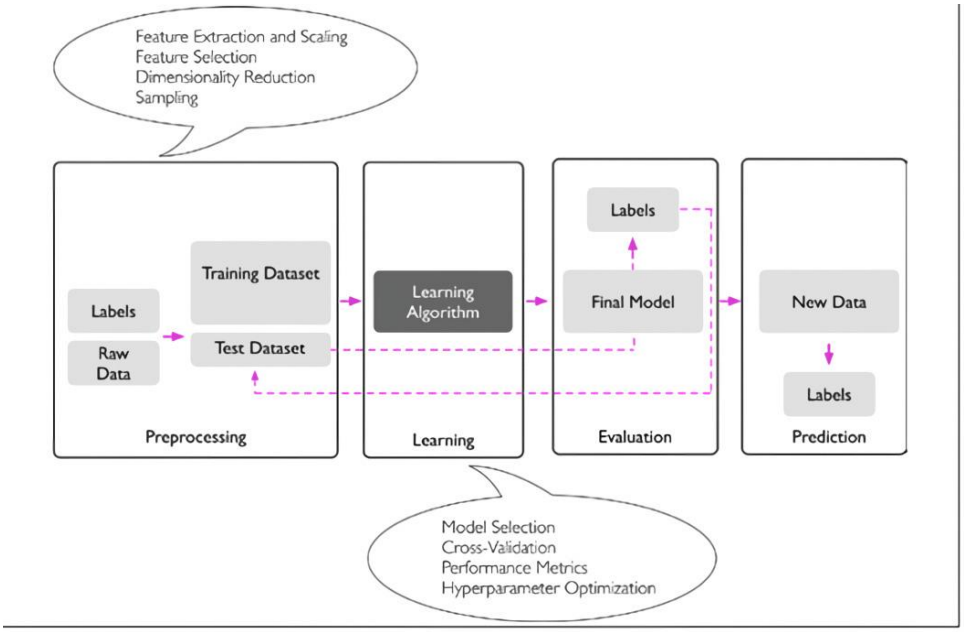
1. Melakukan prediksi sentimen dengan model Logistic Regression dan Random Forest.
2. Melakukan perbandingan kinerja dari masing-masing model dalam memprediksi sentimen.
3. Implementasi web untuk model Logistic Regression dan Random Forest.

Ruang lingkup tugas akhir ini mencakup:

1. Penggunaan dataset review film berbahasa Inggris.
2. Perbandingan dua model: Logistic Regression dan Random Forest.
3. Hasil penelitian berupa perbandingan kinerja kedua model dan antarmuka web dengan model terbaik.
4. Pengguna akan memahami kinerja kedua model dalam memprediksi sentimen.
5. Penelitian ini dapat digunakan sebagai panduan bagi pengembang perangkat lunak yang tertarik dalam pengembangan aplikasi atau sistem analisis sentimen untuk memilih model yang paling sesuai dengan kebutuhan mereka.

II. KAJIAN TEORI

Machine learning adalah metode yang memungkinkan komputer mempelajari dan melakukan tugas secara otomatis melalui algoritma tertentu. Menurut Hairani (2023), proses ini dilakukan dalam dua tahap: pelatihan dan pemanfaatan. Pelatihan melibatkan pemodelan algoritma menggunakan data pelatihan, sedangkan pemanfaatan menggunakan model yang telah dipelajari untuk membuat keputusan dengan data pengujian. Ada dua jenis pembelajaran mesin: supervised learning dan unsupervised learning. Unsupervised learning memproses data tanpa label atau kelas, digunakan dalam visualisasi dan clustering, sedangkan supervised learning menggunakan label untuk klasifikasi.



Gambar 1. Metodologi

Pengumpulan Data Pengumpulan data adalah langkah awal dalam penelitian machine learning. Sumber data bisa berasal dari database internal, API, web scraping, atau dataset publik yang tersedia online.

Preprocessing Data Data mentah sering tidak dalam format yang dibutuhkan untuk algoritma pembelajaran. Langkah penting ini mencakup pembersihan data, seperti menangani nilai yang hilang dengan estimasi atau penghapusan, dan penyandian kategori menggunakan metode seperti One-Hot Encoding dan Label Encoding.

Pembagian Dataset Dataset dibagi menjadi set pelatihan dan pengujian. Set pelatihan digunakan untuk melatih model, sementara set pengujian digunakan untuk mengevaluasi kinerja model.

Tahapan Machine Learning

Pelatihan (Training): Proses pemodelan algoritma menggunakan data pelatihan untuk mempelajari pola atau relasi antara data input dan output (label).

Pemanfaatan: Menggunakan model yang telah dipelajari untuk membuat prediksi atau keputusan berdasarkan data pengujian.

Dalam supervised learning, algoritma mempelajari hubungan antara data input (x) dan label (y) melalui proses klasifikasi. Dalam unsupervised learning, data diproses tanpa label untuk menemukan pola atau grup dalam data. Proses preprocessing data sangat penting untuk memastikan data siap digunakan oleh algoritma machine learning, termasuk menangani nilai yang hilang dan mengubah data kategori menjadi format numerik yang dapat dipahami oleh algoritma.

Preprocessing data, pembagian dataset, dan pemilihan sumber data yang tepat adalah langkah-langkah penting dalam membangun sistem machine learning yang efektif. Implementasi machine learning melibatkan pelatihan model dengan data yang bersih dan terstruktur, serta pengujian model untuk mengevaluasi kinerjanya. Pemahaman yang mendalam tentang metode supervised dan unsupervised learning serta teknik preprocessing data adalah kunci keberhasilan dalam penerapan machine learning.

Logistik Regression

Proses klasifikasi biner atau yang hanya memiliki dua kelas target biasanya menggunakan regresi logistik (logistic regression) [19]. Regresi logistik adalah teknik klasifikasi linier yang sederhana dan mudah digunakan, menurut Pan et al. [20]. Salah satu jenis supervised learning adalah regresi logistik [21]. Tingkat signifikansi statistik masing-masing variabel prediktor selama proses pembelajaran dihitung dengan metode probabilitas. Regresi logistik adalah jenis analisis statistik yang digunakan untuk memodelkan hubungan antara satu atau lebih variabel independen dan variabel dependen biner (dua kategori). Ini adalah bagian dari keluarga regresi yang lebih umum, tetapi dimaksudkan untuk menangani tugas klasifikasi biner seperti memprediksi kejadian atau non-kejadian, diagnosis medis (positif atau negatif), dan lainnya. Beberapa konsep dasar dan langkah-langkah dalam Regresi Logistik adalah sebagai berikut:

Fungsi Log-Odds Reggresi logistik menggunakan fungsi log-odds, juga disebut sebagai logit, untuk menunjukkan hubungan antara variabel independen dan kemungkinan suatu peristiwa terjadi. Fungsi log-odds, juga disebut sebagai logit, didefinisikan sebagai berikut Formula Hitung 2.1

$$\text{log-odds} = \ln\left(\frac{p}{1-p}\right)$$

Gambar 2. Fungsi log-odds

p adalah probabilitas kejadian dan 1−p adalah probabilitas non-kejadian.

Fungsi Sigmoid :Agar dapat memetakan log-odds ke probabilitas, menggunakan fungsi sigmoid (fungsi logistik) berikut formula sigmoid 2.20

$$p(X) = \frac{1}{1 + \exp^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}}$$

Gambar 3. Fungsi Sigmoid

adalah probabilitas kejadia X1,X2,…,Xn adalah variabel indevenden , dan β0,β1,…,βn adalalah parameter model

Pilihan Parameter Model Estimasi parameter model adalah bagian dari proses pelatihan regresi logistik (β) yang mengurangi kesalahan prediksi. Metode optimasi seperti metode gradien descent atau metode Newton-Raphson biasanya digunakan untuk melakukan ini.

Keputusan dan Ambang Keputusan Setelah model dilatih, dapat menetapkan ambang keputusan untuk membuat keputusan klasifikasi. Misalnya, jika kemungkinan, yaitu p(X) lebih besar atau sama dengan 0.5, maka kategori itu dianggap positif; sebaliknya, jika kurang dari 0.5, kategori itu dianggap negatif.

Evaluasi Model Bergantung pada kebutuhan dan karakteristik masalah klasifikasi, evaluasi kinerja model regresi logistik melibatkan metrik seperti akurasi, presisi, recall, skor F1, dan area di bawah kurva ROC (AUC-ROC).Akurasi yang Dianggap Baik: 80-90%: Biasanya dianggap baik untuk banyak aplikasi klasifikasi biner standar; 90 persen atau lebih: Dianggap sangat baik, tetapi waspada terhadap overfitting, terutama pada dataset yang kecil.Faktor yang Mempengaruhi Akurasi: Sederhana dan Interpretasi Mudah: Logistic regression sering digunakan sebagai baseline model; jika model memiliki interpretasi yang mudah, akurasi yang lebih rendah mungkin dapat diterima.Data Linear Hubungan antara fitur dan logit probabilitas harus linear, sehingga logistik regression bekerja dengan baik.

Random Forest (RF) adalah algoritma supervised learning yang dikembangkan oleh Breiman pada tahun 2001. Algoritma ini digunakan untuk menyelesaikan masalah klasifikasi dan regresi dengan menggabungkan beberapa pohon keputusan.

Setiap pohon tumbuh dari sampel bootstrap yang diambil secara acak dari data pelatihan.

Selama pembentukan pohon keputusan, subset variabel dipilih secara acak pada setiap node, dan variabel terbaik digunakan untuk pemisahan.

RF menggabungkan beberapa pohon keputusan (tree predictors) yang masing-masing pohonnya tumbuh dari vektor acak yang diambil secara merata dari semua pohon dalam hutan. Prediksi akhir dihasilkan berdasarkan voting suara terbanyak untuk klasifikasi dan rata-rata untuk regresi. Formula untuk prediksi RF adalah:

III. Analisis dan Rancangan Sistem

A. Metode Mengambil Dataset

Untuk mendapatkan Kumpulan review film IMDB, kunjungi http://ai.stanford.edu/~amaas/data/sentiment/ . Setelah mengunduh kumpulan data,lakukan dekompresi file. Kode opsional dapat dipilih untuk mengunduh dan mengekstrak kumpulan data dijelaskan pada Gambar 4

```
import os
import sys
import tarfile
import time
import urllib.request


source = 'http://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz'
target = 'aclImdb_v1.tar.gz'


def reporthook(count, block_size, total_size):
    global start_time
    if count == 0:
        start_time = time.time()
        return
    duration = time.time() - start_time
    progress_size = int(count * block_size)
    speed = progress_size / (1024.**2 * duration)
    percent = count * block_size * 100. / total_size

    sys.stdout.write("\r%d%% | %d MB | %.2f MB/s | %d sec elapsed" %
                     (percent, progress_size / (1024.**2), speed, duration))
    sys.stdout.flush()


if not os.path.isdir('aclImdb') and not os.path.isfile('aclImdb_v1.tar.gz'):
    urllib.request.urlretrieve(source, target, reporthook)
```

Gambar 4. Import Data Set

Dataset dapat diunduh dari URL yang ditentukan dan diekstrak menggunakan skrip Python.Skrip ini mengunduh file tar.gz dari URL tertentu dan menyimpannya secara lokal.Setelah diunduh, skrip mengekstrak isi file ke dalam direktori yang sesuai.

Import Modul yang diperlukan untuk operasi ini termasuk os untuk fungsionalitas sistem operasi, sys untuk parameter khusus sistem, tarfile untuk bekerja dengan arsip tar, time untuk fungsi-fungsi terkait waktu, dan urllib.request untuk membuka dan membaca URL ,dan Nama File URL sumber dari file tar.gz dan nama file lokal untuk menyimpan file yang

diunduh ditentukan dalam skrip.Progress Bar KustomProgress bar kustom digunakan untuk menampilkan kemajuan unduhan dalam bentuk persentase, ukuran, kecepatan unduh, dan waktu yang telah berlalu.Pengecekan dan Ekstraksi Skrip memeriksa apakah direktori bernama 'aclImdb' dan file bernama 'aclImdb_v1.tar.gz' sudah ada di direktori kerja saat ini.Jika direktori belum ada, skrip mengunduh file dan mengekstraknya ke dalam direktori tersebut menggunakan tarfile.open.Ekstraksi Isi ArsipJika 'aclImdb' belum ada, skrip membuka file arsip tar.gz dan mengekstrak seluruh isinya ke dalam direktori baru yang bernama 'aclImdb'.Hal ini memastikan bahwa data dari arsip dapat diakses dan digunakan dalam proyek.

Dengan langkah-langkah ini, dataset ulasan film IMDb dapat diperoleh dan dipersiapkan untuk analisis lebih lanjut dalam penelitian.

B. Preprocessing Dataset

Dalam penelitian ini, dataset ulasan film IMDb digunakan dan dapat diakses melalui link berikut. Dataset tersebut diunduh, diekstraksi, dan diproses untuk analisis sentimen menggunakan bahasa pemrograman Python. Proses preprocessing data dijelaskan sebagai berikut

```
import pyprind
import pandas as pd
import os

# change the `basepath` to the directory of the
# unzipped movie dataset

basepath = 'aclImdb'

labels = {'pos': 1, 'neg': 0}
pbar = pyprind.ProgBar(50000)
df = pd.DataFrame()
for s in ('test', 'train'):
    for l in ('pos', 'neg'):
        path = os.path.join(basepath, s, l)
        for file in sorted(os.listdir(path)):
            with open(os.path.join(path, file),
                      'r', encoding='utf-8') as infile:
                txt = infile.read()
            df = pd.concat([df,pd.DataFrame([[txt, labels[l]]])], ignore_index=True)
            pbar.update()
```

Gambar 5. Preprocesing

Langkah-langkah Preprocessing Mengunduh dan Mengekstrak Dataset , Dataset diunduh dari URL yang ditentukan dan diekstrak menggunakan skrip Python yang menggunakan modul os, sys, tarfile, time, dan urllib.request.Skrip ini memeriksa apakah direktori 'aclImdb' sudah ada. Jika belum, skrip mengunduh file tar.gz dari URL dan mengekstraknya ke dalam direktori tersebut.Membaca dan Memproses Data Modul pyprind, pandas, dan os digunakan untuk membaca ulasan film dari direktori 'aclImdb' dan menyimpannya ke dalam struktur DataFrame.Label untuk ulasan positif ('pos') dan ulasan negatif ('neg') ditetapkan sebagai 1 dan 0.Objek progress bar pyprind.ProgBar digunakan untuk menunjukkan kemajuan iterasi saat membaca file ulasan.Setiap file ulasan dibaca dan disimpan dalam DataFrame bersama dengan labelnya.Penetapan Nama Kolom Nama kolom DataFrame ditetapkan sebagai 'review' dan 'sentiment' untuk membuat DataFrame lebih deskriptif dan memudahkan referensi kolom.Menampilkan DataFrame Metode df.head() digunakan untuk menampilkan sejumlah baris pertama dari DataFrame, memberikan gambaran cepat tentang struktur dan isi data. Pengacakan Data Modul numpy digunakan untuk mengacak baris dalam DataFrame.Seed acak diatur untuk memastikan hasil yang konsisten saat kode dijalankan beberapa kali.DataFrame diacak menggunakan indeks acak yang dihasilkan oleh np.random.permutation(df.index).Langkah-langkah ini memastikan bahwa dataset ulasan film IMDb dipersiapkan dengan baik untuk analisis sentimen, dengan data yang terorganisir dan diacak untuk menghindari bias dalam model pembelajaran mesin.

C. Cleaning Dataset

Cleaning dataset merupakan langkah penting dalam preprocessing data untuk memastikan data berkualitas tinggi dan siap untuk analisis atau pelatihan model pembelajaran mesin. Berikut adalah ringkasan langkah-langkah yang dilakukan dalam proses cleaning dataset ulasan film IMDb Mengatasi Nilai yang Hilang Mengidentifikasi dan menangani nilai yang hilang dalam dataset. Ini bisa dilakukan dengan menghapus baris atau kolom yang memiliki nilai hilang, atau menggantinya dengan nilai estimasi yang sesuai. Mengubah Data Kategori menjadi Numerik Mengonversi data kategori menjadi format numerik yang dapat dipahami oleh algoritma pembelajaran mesin. Dua metode umum yang digunakan adalah One-Hot Encoding dan Label Encoding. One-Hot Encoding mengubah setiap kategori menjadi vektor biner.Label Encoding mengubah setiap kategori menjadi nilai integer.Menghapus Tanda Baca dan Karakter Khusus Membersihkan teks ulasan dengan menghapus tanda baca, karakter khusus, dan angka yang tidak relevan untuk analisis sentimen.Proses ini membantu

dalam menyederhanakan teks dan mengurangi kompleksitas data.Mengonversi Teks ke Huruf KecilMengubah semua teks ulasan menjadi huruf kecil untuk memastikan konsistensi dan menghindari duplikasi karena perbedaan huruf besar dan kecil. Menghapus Stop Words Menghapus kata-kata umum (stop words) yang tidak memiliki makna signifikan dalam analisis sentimen, seperti "the", "is", "in", dan sebagainya.Menggunakan pustaka seperti NLTK atau spaCy untuk mengidentifikasi dan menghapus stop words.Lemmatization atau Stemming Mengonversi kata-kata ke bentuk dasar mereka menggunakan lemmatization atau stemming.Lemmatization mempertahankan makna kontekstual kata, sementara stemming menghapus akhiran kata untuk mendapatkan bentuk dasar. Tokenization Memecah teks ulasan menjadi kata-kata atau token individu yang dapat dianalisis secara terpisah.Tokenization membantu dalam mempersiapkan data untuk pemrosesan lebih lanjut dalam model pembelajaran mesin.Proses cleaning ini memastikan bahwa dataset ulasan film IMDb bebas dari noise, terstruktur dengan baik, dan siap untuk dianalisis atau digunakan dalam pelatihan model pembelajaran mesin untuk prediksi sentimen. Langkah-langkah ini juga membantu dalam meningkatkan akurasi dan kinerja model dengan menyediakan data yang bersih dan relevan.

D. Processing document into tokens

Tokenisasi dokumen adalah langkah penting dalam preprocessing teks yang memecah teks menjadi unit-unit yang lebih kecil, yang disebut token. Tokenisasi memudahkan analisis dan pemrosesan teks lebih lanjut dalam model pembelajaran mesin. Berikut adalah penjelasan singkat tentang tokenisasi untuk artikel jurnal Proses Tokenisasi Definisi Tokenisasi Tokenisasi adalah proses membagi teks menjadi token-token yang lebih kecil, seperti kata, frasa, atau kalimat. Dalam konteks analisis sentimen, token biasanya merujuk pada kata-kata individual.Langkah-langkah Tokenisasi Penghapusan Karakter Khusus: Menghapus tanda baca, angka, dan karakter khusus yang tidak relevan untuk analisis sentimen.Konversi ke Huruf Kecil: Mengonversi semua teks ke huruf kecil untuk konsistensi dan menghindari perbedaan karena kapitalisasi. Penghapusan Stop Words: Menghapus kata-kata umum (stop words) yang tidak signifikan dalam analisis sentimen, seperti "the", "is", "in", dll. Lemmatization atau Stemming: Mengonversi kata-kata ke bentuk dasar atau akarnya untuk konsistensi (misalnya, "running" menjadi "run").Implementasi Tokenisasi Menggunakan pustaka pemrosesan bahasa alami seperti NLTK, spaCy, atau gensim untuk melakukan tokenisasi.

```
from nltk.stem.porter import PorterStemmer

porter = PorterStemmer()

def tokenizer(text):
    return text.split()


def tokenizer_porter(text):
    return [porter.stem(word) for word in text.split()]
```

Gambar 6. Tokenisasi

Analisis Frekuensi Kata: Memungkinkan analisis frekuensi kata untuk memahami distribusi kata dalam teks.Pembangunan Model: Memfasilitasi pembangunan model pembelajaran mesin dengan menyediakan input yang terstruktur.Pemrosesan Lebih Lanjut: Menyiapkan data untuk langkah-langkah pemrosesan lebih lanjut seperti pembuatan vektor fitur dan penghitungan tf-idf. Tokenisasi adalah langkah mendasar dalam preprocessing teks yang memungkinkan analisis dan pemrosesan lebih lanjut dalam berbagai aplikasi pemrosesan bahasa alami, termasuk analisis sentimen. Dengan membagi teks menjadi token-token yang lebih kecil, kita dapat melakukan analisis yang lebih mendalam dan membangun model pembelajaran mesin yang lebih efektif.

E. Processing document into tokens

Tokenisasi dokumen adalah langkah penting dalam preprocessing teks yang memecah teks menjadi unit-unit yang lebih kecil, yang disebut token. Tokenisasi memudahkan analisis dan pemrosesan teks lebih lanjut dalam model pembelajaran mesin. Berikut adalah penjelasan singkat tentang tokenisasi untuk artikel jurnal Proses Tokenisasi Definisi Tokenisasi Tokenisasi adalah proses membagi teks menjadi token-token yang lebih kecil, seperti kata, frasa, atau kalimat. Dalam konteks analisis sentimen, token biasanya merujuk pada kata-kata individual.Langkah-langkah Tokenisasi Penghapusan Karakter Khusus: Menghapus tanda baca, angka, dan karakter khusus yang idak relevan untuk analisis sentimen.Konversi ke Huruf Kecil: Mengonversi semua teks ke huruf kecil untuk konsistensi dan menghindari perbedaan karena kapitalisasi. Penghapusan Stop Words: Menghapus kata-kata umum (stop words) yang tidak signifikan dalam analisis sentimen, seperti

"the", "is", "in", dll. Lemmatization atau Stemming: Mengonversi kata-kata ke bentuk dasar atau akarnya untuk konsistensi (misalnya, "running" menjadi "run").Implementasi Tokenisasi Menggunakan pustaka pemrosesan bahasa alami seperti NLTK, spaCy, atau gensim untuk melakukan tokenisasi.

```
from nltk.stem.porter import PorterStemmer

porter = PorterStemmer()

def tokenizer(text):
    return text.split()


def tokenizer_porter(text):
    return [porter.stem(word) for word in text.split()]
```

Gambar 7. Tokenisasi

Analisis Frekuensi Kata: Memungkinkan analisis frekuensi kata untuk memahami distribusi kata dalam teks.Pembangunan Model Memfasilitasi pembangunan model pembelajaran mesin dengan menyediakan input yang terstruktur.Pemrosesan Lebih Lanjut: Menyiapkan data untuk langkah-langkah pemrosesan lebih lanjut seperti pembuatan vektor fitur dan penghitungan tf-idf. Tokenisasi adalah langkah mendasar dalam preprocessing teks yang memungkinkan analisis dan pemrosesan lebih lanjut dalam berbagai aplikasi pemrosesan bahasa alami, termasuk analisis sentimen. Dengan membagi teks menjadi token-token yang lebih kecil, kita dapat melakukan analisis yang lebih mendalam dan membangun model pembelajaran mesin yang lebih efektif.

F. Training LogisticRegression

Regresi logistik digunakan untuk klasifikasi biner dengan tujuan memprediksi probabilitas kelas tertentu. Dalam penelitian ini, kami menggunakan regresi logistik untuk memprediksi sentimen ulasan film dari dataset IMDb. Proses pelatihan model ini melibatkan langkah-langkah berikut Persiapan Data Dataset ulasan film IMDb dibagi menjadi set pelatihan dan pengujian.Data di-preproses dengan teknik seperti tokenisasi, penghapusan stop words, dan stemming untuk mengubah teks ulasan menjadi fitur numerik yang dapat digunakan oleh model.Pembentukan Fitur Fitur teks diubah menjadi representasi numerik menggunakan metode seperti Term Frequency-Inverse Document Frequency (TF-IDF).Matriks TF-IDF dari teks ulasan digunakan sebagai input untuk model regresi logistik.Pelatihan Model Model regresi logistik dilatih menggunakan set pelatihan.Selama pelatihan, parameter model dioptimalkan menggunakan metode Maximum Likelihood Estimation (MLE) untuk meminimalkan kesalahan prediksi.Proses optimasi dilakukan dengan algoritma iteratif seperti gradient descent. Evaluasi Model Model yang telah dilatih dievaluasi menggunakan set pengujian untuk mengukur kinerjanya.Metode evaluasi mencakup metrik akurasi, precision, recall, dan F1-score untuk menilai kemampuan model dalam memprediksi sentimen ulasan film.Kurva ROC (Receiver Operating Characteristic) dan AUC (Area Under Curve) juga digunakan untuk mengevaluasi performa klasifikasi biner.Hasil evaluasi menunjukkan bahwa model regresi logistik mampu memberikan prediksi sentimen yang akurat dan dapat diandalkan, dengan interpretasi yang mudah dipahami dari koefisien model.

G. Training Random Forest

Regresi logistik digunakan untuk klasifikasi biner dengan tujuan memprediksi probabilitas kelas tertentu. Dalam penelitian ini, kami menggunakan regresi logistik untuk memprediksi sentimen ulasan film dari dataset IMDb. Proses pelatihan model ini melibatkan langkah-langkah berikut Persiapan Data Dataset ulasan film IMDb dibagi menjadi set pelatihan dan pengujian.Data di-preproses dengan teknik seperti tokenisasi, penghapusan stop words, dan stemming untuk mengubah teks ulasan menjadi fitur numerik yang dapat digunakan oleh model.Pembentukan Fitur Fitur teks diubah menjadi representasi numerik menggunakan metode seperti Term Frequency-Inverse Document Frequency (TF-IDF).Matriks TF-IDF dari teks ulasan digunakan sebagai input untuk model regresi logistik.Pelatihan Model Model regresi logistik dilatih menggunakan set pelatihan.Selama pelatihan, parameter model dioptimalkan menggunakan metode Maximum Likelihood Estimation (MLE) untuk meminimalkan kesalahan prediksi.Proses optimasi dilakukan dengan algoritma iteratif seperti gradient descent. Evaluasi Model Model yang telah dilatih dievaluasi menggunakan set pengujian untuk mengukur kinerjanya.Metode evaluasi mencakup metrik akurasi, precision, recall, dan F1-score untuk menilai kemampuan model dalam memprediksi sentimen ulasan film.Kurva ROC (Receiver Operating Characteristic) dan AUC (Area Under Curve) juga digunakan untuk mengevaluasi performa klasifikasi biner.Hasil evaluasi menunjukkan bahwa

model regresi logistik mampu memberikan prediksi sentimen yang akurat dan dapat diandalkan, dengan interpretasi yang mudah dipahami dari koefisien model.

H. Implementasi Website

memberikan gambaran teknis tentang desain, pengembangan, dan integrasi antarmuka web dengan model analisis sentimen machine learning yang telah dilatih sebelumnya. Antarmuka web ini memungkinkan pengguna memasukkan ulasan film dan mendapatkan analisis sentimen secara real-time.Tujuan dan Penggunaan Kemudahan Akses: Memberikan akses mudah ke hasil analisis sentimen.Pengambilan Keputusan: Membantu pengguna membuat keputusan lebih baik dalam memilih film berdasarkan ulasan pengguna.Proses Pengembangan Desain dan Pengembangan: Meliputi rancangan teknis dan pengembangan antarmuka untuk pengalaman pengguna yang optimal.Fitur Tambahan: Menambahkan fitur yang meningkatkan interaksi dan kemudahan penggunaan antarmuka web.Instruksi Penggunaan Cara Menggunakan: Panduan praktis tentang cara menggunakan antarmuka web untuk analisis sentimen.Aplikasi Hasil Analisis: Penjelasan tentang bagaimana hasil analisis dapat digunakan dalam pengambilan keputusan terkait film.Tujuan dari implementasi ini adalah memberikan alat yang mudah digunakan dan bermanfaat bagi pengguna untuk mengeksplorasi dan memanfaatkan sentimen ulasan film IMDb.

Gambar 8. Halaman Home

Pada Gambar 8 Halaman beranda, juga disebut sebagai halaman home, adalah halaman utama sebuah situs web dan tempat pengunjung masuk.

IV. KESIMPULAN DAN SARAN

Tujuan penelitian ini adalah untuk memprediksi sentimen ulasan film IMDb dengan menggunakan model Logistic Regression dan Random Forest, dan kemudian membandingkan keduanya. Studi ini juga menggunakan antarmuka web untuk kedua model tersebut. Pengantar ini akan mengevaluasi bagaimana tujuan penelitian telah dicapai.Pemilihan Model Kami berhasil menggunakan model Logistic Regression dan Random Forest untuk memprediksi sentimen pada ulasan film IMDb. Kami menemukan bahwa kedua model memberikan hasil yang memuaskan, namun Logistic Regression cenderung lebih unggul dalam memprediksi sentimen pada dataset kami.Perbandingan Kinerja Model Melalui perbandingan kinerja antara Logistic Regression dan Random Forest, kami menemukan bahwa Logistic Regression memiliki keunggulan dalam memprediksi sentimen pada dataset ulasan film IMDb yang kami gunakan. Meskipun Random Forest juga memberikan hasil yang baik, Logistic Regression memberikan akurasi yang sedikit lebih tinggi.Implementasi Web Kami berhasil mengimplementasikan antarmuka web untuk model Logistic Regression dan Random Forest. Antarmuka ini memungkinkan pengguna untuk memasukkan teks ulasan film dan secara langsung mendapatkan prediksi sentimen dari kedua model melalui browser web mereka.

Saran pertama adalah meningkatkan Performa Model Regresi Logistik: Penelitian lebih lanjut harus dilakukan untuk meningkatkan kinerja model Logistic Regression karena, dalam beberapa kasus, model ini memiliki hasil yang lebih baik. Jenis penelitian yang dapat dilakukan termasuk mencoba metode preprocessing data yang lebih canggih, mengubah parameter dengan lebih hati-hati, atau bahkan meneliti model klasifikasi alternatif yang dapat memperbaiki kekurangan model ini.

Saran kedua adalah Pengembangan Antarmuka Pengguna yang Lebih Lanjut Antarmuka pengguna telah dirancang dengan baik, tetapi masih bisa disempurnakan. Pertimbangkan untuk menambahkan fitur tambahan seperti animasi yang

lebih halus, validasi input yang lebih ketat, dan integrasi dengan desain responsif untuk memudahkan penggunaan di berbagai perangkat.

Saran ke-3 adalah uji Coba Lanjutan dengan Pengguna Nyata Sangat penting untuk melakukan uji coba lanjutan dengan pengguna nyata untuk memastikan bahwa aplikasi memenuhi kebutuhan pengguna dengan baik. Dengan mengumpulkan umpan balik langsung dari pengguna, Anda dapat menemukan bagian aplikasi yang perlu diperbaiki dan menyesuaikannya dengan umpan balik tersebut.

Saran terakhir adalah Pemeliharaan dan Pembaruan Berkala Terakhir, pastikan aplikasi diperbarui dan diupdate secara berkala, termasuk pembaruan model klasifikasi dan perbaikan bug, untuk memastikan pengalaman pengguna yang optimal dan kepuasan pengguna yang tinggi.

Dengan demikian, penelitian ini memberikan pemahaman yang lebih baik tentang kinerja model Logistic Regression dan Random Forest dalam memprediksi sentimen ulasan film IMDb, serta menyediakan aplikasi praktis dalam bentuk antarmuka web untuk mengakses model-model tersebut secara mudah dan cepat.

DAFTAR PUSTAKA

[1] S. M. Metev & V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.
[2] Fithratullah, M. (2021). Representasi nilai-nilai keberlanjutan Korea dalam film remake
[3] Pavitha, N., Pungliya, V., Raut, A., Bhonsle, R., Purohit, A., Patel, A., & Shashidhar, R.
[4] (2022). Rekomendasi film dan analisis sentimen menggunakan pembelajaran mesin. M. Wegmuller, J. P. von der Weid, P. Oberson, & N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," *Prosiding ECOC'00*, 2000, paper 11.3.4, p. 109.
[5] Rehman, AU, Malik, AK, Raza, B., & Ali, W. (2019). Model Hibrid CNN-LSTM untuk
[6] (2002) The IEEE website. [Online].Tersedia: http://www.ieee.org/
[7] AM Rahat, A. Kahir dan AKM Masum, "Perbandingan Algoritma Naive Bayes dan SVM berdasarkan
[8] *FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.
[9] Kumar, S., De, K., & Roy, PP (2020). Sistem Rekomendasi Film Menggunakan Analisis
[10] Behera, RK, Jena, M., Rath, SK, & Misra, S. (2021). Co-LSTM: Model LSTM konvolusional untuk
Malviya, S., Tiwari, AK, Srivastava, R., & Tiwari, VK (2020). Teknik Pembelajaran Mesin untuk Analisis Sentimen: Tinjauan. SAMRIDDHI : Jurnal Ilmu Fisika, Teknik
[11] *Maulana, R., Rahayuningsih, PA, Irmayani, W., Saputra, D., & Jayanti, WE (2020).*.