

# TUGAS AKHIR

## ANALISIS SENTIMEN REVIEW FILM UNTUK MENGETAHUI POLA PADA *REVIEW*



Fersylia Oktafianny

NPM: 6181901010

PROGRAM STUDI INFORMATIKA  
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS  
UNIVERSITAS KATOLIK PARAHYANGAN  
2024

**FINAL PROJECT**

**SENTIMENT ANALYSIS OF MOVIE REVIEWS TO IDENTIFY  
PATTERNS IN REVIEWS**



**Fersylia Oktafianny**

**NPM: 6181901010**

**DEPARTMENT OF INFORMATICS  
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES  
PARAHYANGAN CATHOLIC UNIVERSITY  
2024**

# LEMBAR PENGESAHAN

## ANALISIS SENTIMEN REVIEW FILM UNTUK MENGETAHUI POLA PADA *REVIEW*

Fersylia Oktafianny

NPM: 6181901010

Bandung, 24 Juni 2024

Menyetujui,

Pembimbing

Digitally signed  
by Mariskha Tri  
Adithia

Mariskha Tri Adithia, P.D.Eng

Ketua Tim Penguji

Digitally signed  
by Keenan  
Adiwijaya Leman

Keenan Adiwijaya Leman, M.T.

Anggota Tim Penguji

Digitally signed  
by Lionov

Lionov, Ph.D.

Mengetahui,

Ketua Program Studi

Digitally signed  
by Lionov

Lionov, Ph.D.

## PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa tugas akhir dengan judul:

### **ANALISIS SENTIMEN REVIEW FILM UNTUK MENGETAHUI POLA PADA *REVIEW***

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,  
Tanggal 24 Juni 2024



Fersylia Oktafianny  
NPM: 6181901010

## ABSTRAK

Film merupakan salah satu sarana hiburan yang dapat dinikmati oleh setiap orang dari berbagai kalangan, usia maupun jenis kelamin, tak hanya menjadi sarana hiburan terkadang film dapat dijadikan sebagai sarana komunikasi untuk menyalurkan perasaan atau pikiran yang tersirat, maka dari itu film kerap disebut sebagai salah satu media massa. Terdapat beberapa faktor yang membuat seseorang memutuskan untuk menonton film tertentu salah satunya yaitu *review* atau ulasan mengenai film tertentu, beberapa website dapat digunakan untuk membagikan ulasan mengenai film tertentu seperti IMDB, flixster dan *rotten tomatoes*. Situs *rotten tomatoes* bekerja dengan mengumpulkan semua ulasan dari berbagai kritikus, kemudian ulasan tersebut akan ditampilkan dalam bentuk persentase, persentase dalam website *rotten tomatoes* ditunjukkan sebagai proporsi kritikus yang memberikan penilaian positif terhadap suatu film. Terdapat pola yang membuat persentase tersebut semakin tinggi seperti ulasan yang diberikan dengan memiliki sentimen yang positif. Melalui analisis ini, diharapkan para seniman dapat meningkatkan kualitas film dari ulasan yang buruk dan mempertahankan kualitas film dari ulasan yang baik.

Pada penelitian ini, dilakukan beberapa hal yaitu analisis untuk melihat pola *review*, dan membuat model klasifikasi untuk nantinya dapat mengklasifikasi sentimen positif dan negatif, pola yang dihasilkan berupa kata-kata kunci untuk masing-masing sentimen. Dalam penelitian ini pun memanfaatkan proses *web scraping* untuk pengumpulan data dan *text mining* untuk pemrosesan data. Untuk melihat *review* sentimen positif dan negatif dengan melihat berdasarkan nilai yang dilakukan *feature extraction* dengan menggunakan *vector space model* yang menggunakan bobot TF-IDF, dengan melihat nilai IDF yang tinggi. Selanjutnya, untuk menampilkan hasil dari kata-kata yang berhubungan untuk sentimen positif maupun negatif dilakukan visualisasi data dengan menggunakan *barchart* dan *wordcloud*. Setelah dilakukan proses *feature extraction* dengan mengubah data teks menjadi data numerik menggunakan TF-IDF, dilakukan pembuatan model menggunakan tiga model klasifikasi yaitu *decision tree*, *logistic regression* dan *random forest*. Untuk mengukur seberapa baik model yang telah yang telah dibuat, digunakan metode *K-Fold Cross Validation* untuk mengevaluasi kinerja model, serta menggunakan metrik evaluasi seperti *accuracy*, *recall*, *precision*, *f1-Score* dan *specifity*.

Dengan menggunakan tiga percobaan model tersebut didapatkan model klasifikasi dengan *logistic regression* memperoleh hasil nilai rata-rata pada evaluasi model, pada metode *K-Fold Cross Validation* dengan K sebanyak 5 maka model berhasil mengklasifikasi data dengan benar sebanyak 69% pada akurasi. Presisi dan recall masing-masing sebesar 58%. Nilai F1-Score juga mencapai 58%, mencerminkan keseimbangan yang kurang baik untuk mengukur prediksi positif yang benar-benar positif yang disebut presisi dan model dapat mengindikasikan dari total kelas positif yang sebenarnya yang disebut *recall*, walaupun hasil kurang baik namun dari percobaan yang telah dilakukan percobaan ini yang paling baik, dan spesifisitas dimana model cukup efektif dalam mengidentifikasi kelas negatif dengan benar yaitu sebesar 75%, model klasifikasi dibangun menggunakan data *preprocessing* secara unigram. Pada penelitian ini pun dibangun perangkat lunak untuk menampilkan hasil analisis yang telah dilakukan, dan pengguna dapat memasukkan input teks *review* sehingga program dapat mengeluarkan hasil sentimennya.

**Kata-kata kunci:** Film, Sentimen Positif dan Negatif, *Web Scraping*, *Text Mining*, *feature extraction*, *Logistic Regression*

## ABSTRACT

Films are one of the entertainment that can be enjoyed by people from various backgrounds, ages, and genders. Not only serving as a form of entertainment, but sometimes movies can also serve as a means of communication to express underlying feelings or thoughts. Therefore, movies are often referred to as one of the mass media. There are several factors that lead someone to decide to watch a particular movie, one of which is reviews or critiques about a specific film. Several websites can be used to share reviews about specific films such as IMDB, flixster, and rotten tomatoes. Rotten Tomatoes works by collecting all reviews from various critics, and then these reviews are displayed in the form of a percentage, indicating the proportion of critics who give a positive rating to a film. There are patterns that contribute to increasing this percentage, such as reviews with positive sentiment. Through this analysis, it is hoped that filmmakers can improve the quality of films from negative reviews and maintain the quality of films from positive reviews.

In this research, several steps were taken: analyzing patterns in reviews and creating a classification model to identify positive and negative sentiments. The patterns consist of keywords for each sentiment. This research utilized web scraping for data collection and text mining for data processing. To identify positive and negative sentiment reviews, feature extraction was performed using the vector space model with TF-IDF weighting, focusing on high IDF values. Next, data visualization was done using bar charts and word clouds to display words associated with positive and negative sentiments. After converting text data into numerical data using TF-IDF, three classification models were built: decision tree, logistic regression, and random Forest. To evaluate the performance of the models, K-Fold Cross Validation was used, along with evaluation metrics such as accuracy, recall, precision, f1-Score, and specificity.

Using three model experiments, the classification model with Logistic Regression achieved an average performance evaluation, utilizing the K-Fold Cross Validation method with K=5. The model correctly classified the data with an accuracy of 69%, with precision and recall each at 58%. The F1-Score also reached 58%, indicating a less balanced measure of true positive predictions (precision) and the model's ability to indicate the actual positive class (recall). Despite the less-than-ideal results, this experiment was the best among those conducted. The specificity was 75%, indicating the model's effectiveness in correctly identifying negative classes. The classification model was built using unigram data preprocessing. Additionally, software was developed in this study to display the analysis results, allowing users to input review texts and receive sentiment analysis outcomes.

**Keywords:** Film, Positive and Negative Sentiments, Web Scraping, Text Mining, feature extraction, Logistic Regression

*Tugas Akhir ini dipersembahkan untuk mama, papa serta keluarga  
dan teman terdekat, tak lupa untuk pribadi saya dan semoga  
bermanfaat bagi masyarakat ...*

## KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa atas segala rahmat dan kasih karuniaNya sehingga penulis diberikan hikmat dan kekuatan untuk menyelesaikan tugas akhir dengan judul “Analisis Sentimen *review* Film Untuk Mengetahui Pola Pada *review*” dengan baik. Dalam kesempatan ini, penulis ingin menyampaikan rasa terima kasih yang tulus kepada :

1. Segala puji syukur, hormat, dan kemuliaan bagi-Nya saya berikan kepada Tuhan Yesus Kristus. Atas anugerah, berkat, dan hikmat bagi penulis, sehingga penulis dapat menyelesaikan tugas akhir ini dengan baik.
2. Kedua orang tua terkasih, Papa Ferry Irwanto, S.Th dan Mama Ella Nurlia, yang telah memberikan dukungan dan menyemangati penulis tanpa henti dalam menyusun tugas akhir ini. Tak lupa keluarga *F-Squad* yaitu: Ferlix Irawan, S.E, Feni, Fernando Indrawan, Philia Shelby Irawan, dan Fhaldyo Eugene Irawan, terima kasih atas kasih sayang, perhatian, dukungan, dan segala doa yang diberikan. Tanpa bimbingan dan motivasi dari mereka, perjalanan ini tak terasa berarti.
3. Ibu Natalia, S.Si, M.Si selaku pembimbing saya, yang telah memberikan arahan dan bimbingan yang sangat berharga, sehingga penulis dapat menyelesaikan tugas akhir ini dengan baik. Serta kepada Bapak Keenan Adiwijaya Leman, S.T, M.T dan Bapak Lionov, PhD. Selaku dosen penguji yang telah memberikan masukan konstruktif bgai penulis dalam penyusunan tugas akhir ini.
4. Serta, kepada Devin Jonathan, S.T selaku orang terkasih, yang telah mendampingi, memberikan semangat, memberikan dukungan, dan memberikan saran kepada penulis yang sangat membangun dalam penyusunan tugas akhir ini tanpa henti. Serta kepada Andrea Nathali Widayat dan Fernando Indrawan pun terima kasih atas dukungannya dan tak lelah untuk menyemangati penulis.
5. Premananda Setyo, S.Kom, ka Indra Permana Sugianto, S.Kom, William Kurniawan, S.T, Otto Nathanael, S.T. Yang telah membantu dan mendukung penulis. Kepada teman-teman seperjuangan IF Unpar 2019 serta “*wombat lover*”, terima kasih atas kebersamaan, dukungan, dan kehangatan yang telah memberikan semangat dalam menyelesaikan studi ini.
6. Terakhir namun tidak kalah pentingnya, penulis mengucapkan terima kasih kepada diri sendiri, terima kasih telah percaya akan diri sendiri, terima kasih atas kerja keras selama ini, dan terima kasih untuk tidak menyerah.

Penulis pun mengucapkan terima kasih kepada seluruh teman dan saudara penulis yang tidak dapat disebutkan satu per satu. Akhir kata semoga Tuhan Yang Maha Esa membalas semua kebaikan kepada seluruh pihak yang telah memberikan bantuan dan dukungan. Penulis menyadari bahwa tugas akhir ini masih jauh dari kata sempurna, untuk itu penulis memohon maaf jika terdapat kesalahan dan penulis berharap tugas akhir ini dapat memberikan manfaat, dan menambahkan ilmu bagi pembacanya.

Bandung, Juni 2024

Penulis

# DAFTAR ISI

<b>KATA PENGANTAR</b>	<b>xv</b>
<b>DAFTAR ISI</b>	<b>xvii</b>
<b>DAFTAR GAMBAR</b>	<b>xix</b>
<b>1 PENDAHULUAN</b>	<b>1</b>
1.1 Latar Belakang . . . . .	1
1.2 Rumusan Masalah . . . . .	3
1.3 Tujuan . . . . .	3
1.4 Batasan Masalah . . . . .	4
1.5 Metodologi . . . . .	4
1.6 Sistematika Pembahasan . . . . .	4
<b>2 LANDASAN TEORI</b>	<b>7</b>
2.1 <i>Web Scraping</i> . . . . .	7
2.1.1 <i>BeautifulSoup</i> [1] . . . . .	7
2.1.2 <i>Selenium</i> . . . . .	9
2.2 <i>Text Mining</i> . . . . .	10
2.2.1 <i>Tokenizing</i> [2] . . . . .	10
2.2.2 <i>N-gram</i> . . . . .	10
2.2.3 <i>Filtering</i> [3] . . . . .	11
2.2.4 <i>Lematisasi</i> [2] . . . . .	11
2.2.5 <i>Stop Word Removal</i> [4] . . . . .	11
2.2.6 <i>VSM (Vector Space Model)</i> [5] . . . . .	11
2.2.7 <i>TF-IDF</i> [6] . . . . .	12
2.2.8 <i>PosTagging</i> [7] . . . . .	13
2.3 <i>Klasifikasi</i> [8] . . . . .	14
2.3.1 <i>Decision Tree</i> . . . . .	14
2.3.2 <i>Logistic Regression</i> . . . . .	16
2.3.3 <i>Random Forest</i> . . . . .	18
2.4 <i>K-Fold Cross Validation</i> . . . . .	19
2.5 <i>Teknik Evaluasi</i> . . . . .	19
2.5.1 <i>Confusion Matrix</i> [8] . . . . .	20
2.5.2 <i>Precision, Recal, Specifity, F1-Score</i> [8] . . . . .	20
<b>3 PENGUMPULAN, EKSPLORASI DAN PENYIAPAN DATA</b>	<b>23</b>
3.1 <i>Pengumpulan dan Penarikan Data</i> . . . . .	23
3.2 <i>Penyiapan Data dan Eksplorasi Data</i> . . . . .	25
3.2.1 <i>Eksplorasi dengan Menggunakan Data Kecil</i> . . . . .	25
3.2.2 <i>Penyiapan dan Pembersihan Data</i> . . . . .	34
3.2.3 <i>Eksplorasi Data</i> . . . . .	37
3.3 <i>Analisa Kata Kunci</i> . . . . .	49

3.3.1	Analisa Kata Kunci Unigram . . . . .	49
3.3.2	Analisa Kata Kunci Bigram . . . . .	53
3.3.3	Analisa Menggunakan PosTag . . . . .	56
<b>4</b>	<b>ANALISIS DATA DAN PENGUJIAN MODEL</b>	<b>59</b>
4.1	Penjelasan Dataset . . . . .	59
4.1.1	Dataset Preprocessing . . . . .	59
4.1.2	Dataset Postag . . . . .	60
4.2	Pembuatan <i>WordCloud</i> . . . . .	60
4.3	Perhitungan Manual Untuk Model Klasifikasi . . . . .	64
4.3.1	Perhitungan manual untuk <i>Decision Tree</i> . . . . .	64
4.3.2	Perhitungan manual untuk <i>Random Forest</i> . . . . .	68
4.3.3	Perhitungan manual untuk <i>Logistic Regression</i> . . . . .	70
4.4	Pembuatan dan Pengujian Model . . . . .	74
4.4.1	<i>Decision Tree</i> . . . . .	74
4.4.2	<i>Logistic Regression</i> . . . . .	82
4.4.3	<i>Random Forest</i> . . . . .	89
<b>5</b>	<b>PERANCANGAN <i>Website</i> DAN IMPLEMENTASINYA</b>	<b>99</b>
5.1	Fitur Perangkat Lunak . . . . .	99
5.2	Diagram <i>Use Case</i> . . . . .	99
5.3	Implementasi Perangkat Lunak Berbasis <i>Website</i> . . . . .	100
<b>6</b>	<b>KESIMPULAN DAN SARAN</b>	<b>103</b>
6.1	Kesimpulan . . . . .	103
6.2	Saran . . . . .	104
	<b>DAFTAR REFERENSI</b>	<b>105</b>
	<b>A KODE PROGRAM</b>	<b>107</b>
	<b>B HASIL EKSPERIMEN</b>	<b>125</b>

## DAFTAR GAMBAR

1.1	Contoh Ulasan Film <i>The Last Of Us</i> dalam situs <i>Rotten Tomatoes</i> . . . . .	2
1.2	Contoh Ulasan Film <i>The Last Of Us</i> dalam situs <i>Rotten Tomatoes</i> dengan Predikat <i>Fresh and Rotten</i> . . . . .	2
2.1	Ilustrasi Pohon dalam HTML . . . . .	8
2.2	<i>Penn Treebank Part of Speech Tags</i> . . . . .	14
2.3	Contoh <i>Decision Tree</i> . . . . .	15
2.4	Ilustrasi Gambar <i>K-Fold Cross Validation</i> . . . . .	19
2.5	<i>Confusion matrix</i> <a href="#">Sumber Ilustrasi</a> . . . . .	20
3.1	Hasil Kata dari Perhitungan TF-IDF Menggunakan Python . . . . .	31
3.2	Hasil Kata dari Perhitungan TF-IDF Secara Sintaks Python dengan Bigram . . . . .	33
3.3	Dokumen Sebelum Melakukan Proses Penghapusan Tanda Baca . . . . .	34
3.4	Dokumen yang telah melakukan proses menghapus tanda baca . . . . .	34
3.5	Hasil tokenisasi setelah melakukan tahap penghapusan <i>stopwords</i> . . . . .	35
3.6	Hasil Penggabungan Tokenisasi Setelah Melakukan <i>Preprocessing</i> . . . . .	36
3.7	Dataset Menggunakan Atribut yang Dipilih . . . . .	37
3.8	Hasil nilai IDF dengan kata yang jarang tampil . . . . .	39
3.9	Hasil nilai IDF dengan kata yang sering tampil . . . . .	39
3.10	Hasil Nilai TF-IDF 10 Tertinggi dan 10 Terendah . . . . .	40
3.11	Jumlah dataset yang ada pada kelas sentimen positif dan negatif . . . . .	41
3.12	Proporsi dari sentimen positif dan negatif . . . . .	41
3.13	Hasil nilai IDF dengan kata yang jarang tampil pada dataset ulasan positif . . . . .	42
3.14	Hasil nilai IDF dengan kata yang sering tampil pada dataset ulasan positif . . . . .	42
3.15	Hasil Nilai TF-IDF 10 Tertinggi Pada Data Berlabel Positif . . . . .	43
3.16	Hasil Nilai TF-IDF 10 Terendah Pada Data Berlabel Positif . . . . .	43
3.17	Hasil nilai IDF dengan kata yang jarang tampil pada dataset ulasan negatif . . . . .	44
3.18	Hasil nilai IDF dengan kata yang sering tampil pada dataset ulasan negatif . . . . .	44
3.19	Hasil Nilai TF-IDF 10 Tertinggi Pada Data Berlabel Negatif . . . . .	44
3.20	Hasil Nilai TF-IDF 10 Terendah Pada Data Berlabel Negatif . . . . .	45
3.21	Jumlah Kemunculan Kata Pada Dataset Berlabelan Positif . . . . .	45
3.22	Jumlah Kemunculan Kata Pada Dataset Berlabelan Negatif . . . . .	46
3.23	Hasil Nilai TF-IDF 10 Tertinggi dengan Bigram . . . . .	46
3.24	Hasil Nilai TF-IDF 10 Terendah dengan Bigram . . . . .	47
3.25	Hasil Nilai TF-IDF 10 Tertinggi dengan Bigram pada Label Positif . . . . .	47
3.26	Hasil Nilai TF-IDF 10 Tertinggi dengan Bigram pada Label Negatif . . . . .	48
3.27	Hasil Nilai TF-IDF 10 Terendah dengan Bigram pada Label Positif . . . . .	48
3.28	Hasil Nilai TF-IDF 10 Terendah dengan Bigram pada Label Negatif . . . . .	48
3.29	Histogram Frekuensi PosTag Sebelum Dianalisis . . . . .	56
3.30	Histogram Frekuensi PosTag Setelah Dianalisis . . . . .	57
3.31	Hasil 10 Nilai IDF Terendah dari Data PosTag . . . . .	57
3.32	Hasil 10 Nilai IDF Tertinggi dari Data PosTag . . . . .	57

4.1	Hasil <i>Wordcloud</i> dari Data <i>Preprocessing</i> Positif Unigram . . . . .	61
4.2	Hasil <i>Wordcloud</i> dari Data <i>Preprocessing</i> Negatif Unigram . . . . .	62
4.3	Hasil <i>Wordcloud</i> dari Data <i>Preprocessing</i> Positif Bigram . . . . .	62
4.4	Hasil <i>Wordcloud</i> dari Data <i>Preprocessing</i> Negatif Bigram . . . . .	63
4.5	Hasil <i>Wordcloud</i> dari Data <i>Postag</i> Positif Unigram . . . . .	63
4.6	Hasil <i>Wordcloud</i> dari Data <i>Postag</i> Negatif Unigram . . . . .	64
4.7	Hasil Visualisasi <i>Decision Tree</i> . . . . .	67
4.8	Hasil Visualisasi <i>Decision Tree</i> Untuk <i>Random Forest</i> Data Subset Pertama . . . . .	68
4.9	Hasil Visualisasi <i>Decision Tree</i> Untuk <i>Random Forest</i> Data Subset Kedua . . . . .	69
4.10	Hasil Visualisasi <i>Decision Tree</i> Untuk <i>Random Forest</i> Data Subset Ketiga . . . . .	69
4.11	<i>Confusion Matrix</i> Untuk <i>Fold</i> Pertama . . . . .	75
4.12	<i>Confusion Matrix</i> Untuk <i>Fold</i> Kedua . . . . .	76
4.13	<i>Confusion Matrix</i> Untuk <i>Fold</i> Ketiga . . . . .	76
4.14	<i>Confusion Matrix</i> Untuk <i>Fold</i> Keempat . . . . .	77
4.15	<i>Confusion Matrix</i> Untuk <i>Fold</i> Kelima . . . . .	78
4.16	<i>Confusion Matrix Decision Tree</i> Untuk <i>Fold</i> Pertama Data <i>Postag</i> . . . . .	79
4.17	<i>Confusion Matrix Decision Tree</i> Untuk <i>Fold</i> Kedua Data <i>Postag</i> . . . . .	79
4.18	<i>Confusion Matrix Decision Tree</i> Untuk <i>Fold</i> Ketiga Data <i>Postag</i> . . . . .	80
4.19	<i>Confusion Matrix Decision Tree</i> Untuk <i>Fold</i> Keempat Data <i>Postag</i> . . . . .	81
4.20	<i>Confusion Matrix Decision Tree</i> Untuk <i>Fold</i> Kelima Data <i>Postag</i> . . . . .	81
4.21	<i>Confusion Matrix</i> Logistik Regresi Untuk <i>Fold</i> Pertama . . . . .	83
4.22	<i>Confusion Matrix</i> Logistik Regresi Untuk <i>Fold</i> Kedua . . . . .	83
4.23	<i>Confusion Matrix</i> Logistik Regresi Untuk <i>Fold</i> Ketiga . . . . .	84
4.24	<i>Confusion Matrix</i> Logistik Regresi Untuk <i>Fold</i> Keempat . . . . .	84
4.25	<i>Confusion Matrix</i> Logistik Regresi Untuk <i>Fold</i> Kelima . . . . .	85
4.26	<i>Confusion Matrix</i> Logistik Regresi Untuk <i>Fold</i> Pertama Data <i>Postag</i> . . . . .	86
4.27	<i>Confusion Matrix</i> Logistik Regresi Untuk <i>Fold</i> Kedua Data <i>Postag</i> . . . . .	87
4.28	<i>Confusion Matrix</i> Logistik Regresi Untuk <i>Fold</i> Ketiga Data <i>Postag</i> . . . . .	87
4.29	<i>Confusion Matrix</i> Logistik Regresi Untuk <i>Fold</i> Keempat Data <i>Postag</i> . . . . .	88
4.30	<i>Confusion Matrix</i> Logistik Regresi Untuk <i>Fold</i> Kelima Data <i>Postag</i> . . . . .	89
4.31	<i>Confusion Matrix Random Forest</i> Untuk <i>Fold</i> Pertama . . . . .	90
4.32	<i>Confusion Matrix Random Forest</i> Untuk <i>Fold</i> Kedua . . . . .	91
4.33	<i>Confusion Matrix Random Forest</i> Untuk <i>Fold</i> Ketiga . . . . .	91
4.34	<i>Confusion Matrix Random Forest</i> Untuk <i>Fold</i> Keempat . . . . .	92
4.35	<i>Confusion Matrix Random Forest</i> Untuk <i>Fold</i> Kelima . . . . .	92
4.36	<i>Confusion Matrix Random Forest</i> Untuk <i>Fold</i> Pertama Data <i>Postag</i> . . . . .	94
4.37	<i>Confusion Matrix Random Forest</i> Untuk <i>Fold</i> Kedua Data <i>Postag</i> . . . . .	94
4.38	<i>Confusion Matrix Random Forest</i> Untuk <i>Fold</i> Ketiga Data <i>Postag</i> . . . . .	95
4.39	<i>Confusion Matrix Random Forest</i> Untuk <i>Fold</i> Keempat Data <i>Postag</i> . . . . .	96
4.40	<i>Confusion Matrix Random Forest</i> Untuk <i>Fold</i> Kelima Data <i>Postag</i> . . . . .	96
5.1	Diagram <i>Use Case</i> Perangkat Lunak Berbasis <i>Website</i> untuk Sentimen Analisis <i>Review</i> Film . . . . .	99
5.2	Halaman <i>dashboard</i> Penjelasan dataset . . . . .	100
5.3	Halaman <i>dashboard</i> Hasil <i>Wordcloud</i> . . . . .	101
5.4	Halaman <i>dashboard</i> Hasil Pengujian Model . . . . .	101
5.5	Halaman <i>input review</i> teks . . . . .	101
5.6	Pengguna Telah Memasukkan <i>Review</i> Teks . . . . .	102
5.7	Hasil Sentimen <i>Review</i> Teks . . . . .	102
B.1	Hasil <i>Wordcloud</i> dari Data <i>Preprocessing</i> Positif Unigram . . . . .	127
B.2	Hasil <i>Wordcloud</i> dari Data <i>Preprocessing</i> Negatif Unigram . . . . .	128

B.3	Hasil <i>Wordcloud</i> dari Data <i>Preprocessing</i> Positif Bigram . . . . .	128
B.4	Hasil <i>Wordcloud</i> dari Data <i>Preprocessing</i> Negatif . . . . .	128
B.5	Hasil <i>Wordcloud</i> dari Data <i>Postag</i> Positif Unigram . . . . .	129
B.6	Hasil <i>Wordcloud</i> dari Data <i>Postag</i> Negatif Unigram . . . . .	129
B.7	Hasil <i>Wordcloud</i> dari Data <i>Postag</i> Positif Bigram . . . . .	129
B.8	Hasil <i>Wordcloud</i> dari Data <i>Postag</i> Negatif Bigram . . . . .	130

# BAB 1

## PENDAHULUAN

### 1.1 Latar Belakang

Film atau *movie* merupakan salah satu sarana hiburan yang dapat dinikmati oleh setiap orang dari berbagai kalangan, usia maupun jenis kelamin, tak hanya menjadi sarana hiburan terkadang film dapat dijadikan sebagai sarana komunikasi untuk menyalurkan perasaan atau pikiran yang tersirat, maka dari itu film kerap disebut sebagai salah satu media massa. Tak sedikit orang kerap menghabiskan waktu hanya dengan menonton film, maka dari itu sebuah film dapat dikatakan sukses jika film tersebut banyak diminati oleh masyarakat.

Terdapat beberapa faktor yang membuat seseorang memutuskan untuk menonton film tersebut salah satunya yaitu *review* atau ulasan mengenai film tersebut. Ulasan tak hanya terdapat di dalam film saja, ulasan dapat digunakan untuk mengulas mengenai suatu buku, jurnal, artikel dan karya-karya lainnya, suatu ulasan diberikan bertujuan sebagai pemberi informasi, maka dari itu tak sedikit orang sebelum menonton film akan melihat ulasan mengenai film tersebut sehingga masyarakat dapat memiliki bayangan. Pada ulasan film biasanya penonton yang telah menonton film tersebut akan memberikan penilaian hingga suatu kritik pada film tersebut. Dengan mengulas suatu karya dapat meningkatkan rasa kepercayaan akan suatu karya tersebut.

Ulasan dapat disampaikan di berbagai media sosial yang disediakan yaitu pada twitter, instagram, dan facebook. Beberapa website dapat digunakan untuk membagikan ulasan mengenai film tersebut seperti IMDB, flixster dan *rotten tomatoes*. Situs *rotten tomatoes* bekerja dengan mengumpulkan semua ulasan dari berbagai kritikus, kemudian ulasan tersebut akan ditampilkan dalam bentuk persentase, persentase dalam *website rotten tomatoes* ditunjukkan sebagai proporsi kritikus yang memberikan penilaian positif terhadap suatu film. Namun, tak hanya kritikus yang dapat memberikan *review*, masyarakat yang menonton pun dapat menuliskan ulasannya. Pada Gambar 1.1 ditunjukkan penilaian yang berasal dari kritikus diberikan *icon* tomat dan penilaian dari penonton diberikan ikon *popcorn*

Menurut Jeff Voris selaku *Vice President* dari *Rotten Tomatoes* tujuan dari situs tersebut yaitu menyediakan sarana untuk para penggemar agar dapat membantu keputusan dalam menonton. Pada situs *Rotten Tomatoes* terdapat persentase yang menentukan baik atau buruknya sebuah film, persentase tersebut memberikan predikat tertentu seperti *rotten* yang menunjukkan persentase terburuk pada Gambar 1.2. Untuk mendapatkan predikat "*Certified Fresh*", sebuah film harus mendapatkan persentase sebesar 75% atau lebih. Selain itu film tersebut harus memiliki setidaknya 40 ulasan yang memiliki nilai positif atau *fresh* pada Gambar 1.2. Selain itu, setidaknya 5 ulasan berasal dari *Top critics*<sup>1</sup>. Contohnya, seperti pada Gambar 1.1 film tersebut mendapatkan predikat "*Certified Fresh*" dikarenakan persentase telah mencapai 96%.

Dari berbagai data *review* film yang diberikan oleh kritikus dan penonton, terdapat pola yang membuat persentase tersebut semakin tinggi seperti ulasan yang diberikan memiliki sentimen

---

<sup>1</sup>"*Rotten Tomatoes, explained*", Alissa Wilkinson, 2018, <https://www.vox.com/culture/2017/8/31/16107948/rotten-tomatoes-score-get-their-ratings-top-critics-certified-fresh-aggregate-mean>, diakses pada 30 Januari 2023

**Rotten Tomatoes** Search movies, TV, actors, more... MOVIES TV SHOWS

TRENDING ON RT The Last of Us Creed III Cocaine Bear Scream VI

HOME > THE LAST OF US > SEASON 1

**THE LAST OF US** (2023)

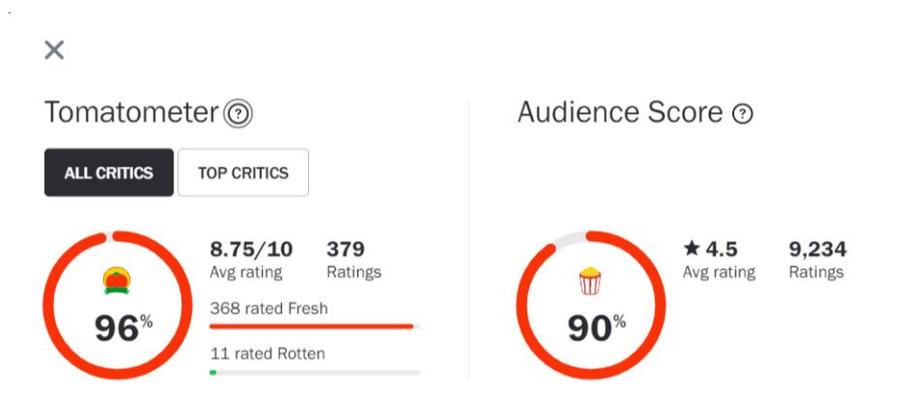
**SEASON 1**  
THE LAST OF US

**Critics Consensus**  
Retaining the most addictive aspects of its beloved source material while digging deeper into the story, *The Last of Us* is binge-worthy TV that ranks among the all-time greatest video game adaptations.

**FRESH** **96%** **90%**  
TOMATOMETER AUDIENCE SCORE  
Critic Ratings: 379 User Ratings: 9234

WANT TO SEE EPISODE LIST SEE SCORE DETAILS

Gambar 1.1: Contoh Ulasan Film *The Last Of Us* dalam situs *Rotten Tomatoes*



Gambar 1.2: Contoh Ulasan Film *The Last Of Us* dalam situs *Rotten Tomatoes* dengan Predikat *Fresh and Rotten*

yang positif<sup>2</sup>, tetapi tidak sedikit pula seseorang memberikan ulasan yang memiliki arti netral yaitu tidak memiliki arti positif maupun arti negatif. Pada analisis ini dilakukan dengan menggunakan sentimen positif dan negatif saja, analisis ini dilakukan dengan melihat nilai berdasarkan nilai IDF yang tinggi, namun dilihat pula kata kunci dengan nilai IDF yang rendah. Dengan menganalisis ulasan maka dapat mengetahui kata kunci yang mempengaruhi baik atau buruknya suatu film. Serta diharapkan para seniman dapat meningkatkan kualitas film dari ulasan yang buruk dan mempertahankan kualitas film dari ulasan yang baik. Seperti contoh jika pada *wordcloud* yang memiliki sentimen positif terdapat kata “*best performance*” hal tersebut dapat dijadikan *insight* yang berguna yaitu “*performance*” pada film sudah baik, begitu pula dengan sentimen negatif.

Pada penelitian ini akan dilakukan beberapa proses. Proses pertama yaitu mengambil dataset yang akan digunakan dengan metode *web scraping* untuk mengambil ulasan pada *website*, pada pengambilan data dilakukan menggunakan *web scraping* selenium. Proses selanjutnya yaitu setelah data berhasil diambil dan dikumpulkan data tersebut akan dilakukan tahap *preprocessing*, data “kotor” yang digunakan akan dipersiapkan terlebih dahulu dengan cara menormalisasikan data seperti melakukan mengubah huruf besar menjadi huruf kecil hal ini bertujuan agar data konsisten serta mengurangi variasi kata, selain itu dilakukan juga menghapus karakter khusus dan tanda baca, tahapan ini dilakukan untuk membantu mengurangi dimensi fitur yang akan dilakukan pada langkah selanjutnya dengan menggunakan TF-IDF. Tahapan yang dilakukan dalam *preprocessing* juga meliputi pembuatan tokenisasi, menghapus *stopwords*, lematisasi, dan penggabungan kalimat.

Proses selanjutnya pada penelitian ini yaitu melakukan *feature extraction* dengan menggunakan metode TF-IDF, data yang digunakan untuk melakukan *feature extraction* yaitu menggunakan data *preprocessing* dan *postag*. Data yang telah diubah menjadi numerik digunakan untuk menganalisis kata kunci yang mempengaruhi sentimen serta menjadi masukan untuk model klasifikasi. Digunakan nilai *inverse document frequency* (IDF) pada analisis kata kunci untuk menilai kepentingan suatu kata pada data *preprocessing* dan *postag*. Kata kunci yang telah dianalisis akan dilakukan pembuatan visualisasi menggunakan *wordcloud*.

Fitur-fitur yang telah dianalisis dengan menggunakan *feature extraction* akan digunakan untuk membuat model klasifikasi *decision tree*, *logistic regression*, dan *random forest*. Proses selanjutnya yaitu menguji model klasifikasi yang telah dibangun dengan tujuan untuk mengukur performa model. Pada proses terakhir yaitu mengimplementasikan hasil analisis pada perangkat lunak berbasis *website* di mana pengguna dapat melihat hasil dari analisis pada penelitian ini yaitu melihat penjelasan dataset yang digunakan serta melihat visualisasi *wordcloud* untuk kata kunci positif dan negatif secara *preprocessing* dan *postag*. Pengguna dapat memasukan sebuah *review* nantinya pengguna akan mendapatkan hasil sentimen dari sebuah *review* yang diberikan oleh pengguna.

## 1.2 Rumusan Masalah

Beberapa rumusan Masalah yang muncul berdasarkan latar belakang adalah sebagai berikut :

1. Bagaimana cara melakukan penarikan data menggunakan *web scraping*?
2. Bagaimana cara mengekstraksi fitur pada *review* film?
3. Bagaimana cara melakukan pembuatan model klasifikasi untuk menentukan sentimen positif dan negatif pada *review* film?
4. Bagaimana cara mengevaluasi model klasifikasi pada ulasan film?
5. Bagaimana cara mengimplementasikan hasil analisis kata kunci *wordcloud* dan pengujian model klasifikasi pada perangkat lunak berbasis web?

## 1.3 Tujuan

Adapun beberapa tujuan yang akan dicapai melalui penelitian ini adalah sebagai berikut :

<sup>2</sup>: “About Rotten Tomatoes”, <https://www.rottentomatoes.com/about> , diakses pada 26 Desember 2023

1. Mempelajari cara *web scraping* dan melakukan penarikan data menggunakan *web scraping*.
2. Mempelajari cara mengekstraksi sentimen *review* film dan melakukan ekstraksi sentimen *review* film.
3. Melakukan pembuatan model klasifikasi untuk menentukan sentimen positif dan negatif pada *review* film.
4. Mempelajari teknik algoritma untuk mengevaluasi model klasifikasi dan melakukan evaluasi model.
5. Membuat perangkat lunak berbasis *web* untuk menampilkan hasil analisis kata kunci *wordcloud* dan hasil pengujian model klasifikasi.

## 1.4 Batasan Masalah

Pada analisa ini hanya akan dibahas mengenai masalah-masalah berikut:

1. Penelitian ini hanya berfokus dengan menggunakan sentimen positif dan negatif sebagai dasar analisis kata kunci serta pembuatan model.
2. Pada penelitian ini, data yang diambil hanya dari *website Rotten Tomatoes* saja.

## 1.5 Metodologi

Metodologi penelitian yang akan digunakan dalam penelitian ini adalah sebagai berikut:

1. Melakukan studi literatur untuk menambah wawasan mengenai penelitian ini, dan juga digunakan untuk mengetahui beberapa teori yang akan dipakai dalam penelitian ini, seperti teori klasifikasi, dan *web scraping*.
2. Penarikan dan pengumpulan data dengan menggunakan metode *webscraping* Selenium.
3. Eksplorasi dan pembersihan data guna mempersiapkan dataset sehingga dapat dilakukan analisis lebih lanjut. Eksplorasi data meliputi pembuatan *bag of words*, TFIDF, dan *postagging*, serta melihat proporsi dari sentimen positif dan negatif.
4. Pembuatan model klasifikasi dan Pengujian dengan model *decision tree*, *logistic regression*, dan *random forest* dengan menggunakan data DTM(*Document Term Matrix*) secara TFIDF. Lalu melakukan pengujian dengan menggunakan *K-fold cross validation* untuk mengetahui model mana yang terbaik serta melakukan teknik evaluasi menggunakan *precision*, *recall*, dan *specifity*.
5. Perancangan dan pembuatan perangkat lunak untuk menampilkan hasil analisis sentimen. Perangkat lunak pun dapat menerima input berupa teks dan akan mengeluarkan sebuah sentimen.
6. Penulisan dokumen skripsi mengenai penelitian yang telah dilakukan.

## 1.6 Sistematika Pembahasan

Sistematika pembahasan laporan penelitian dibagi kedalam enam bab adalah sebagai berikut:

1. Bab 1 : Pendahuluan  
Pada bab ini membahas mengenai latar belakang permasalahan yang ingin diteliti, serta pada bab ini pun membahas mengenai rumusan masalah, tujuan yang akan dicapai pada penelitian ini serta menentukan batasan masalah dari penelitian ini.
2. Bab 2 : landasan Teori  
Pada bab ini membahas mengenai hasil studi literatur yang nantinya akan menjadi dasar teori dalam penelitian ini. Pada bab landasan teori membahas mengenai *text mining*, model klasifikasi, *feature exctraction*, dan teknik evaluasi.
3. Bab 3 : Pengumpulan, Eksplorasi Dan Penyiapan Data  
Pada bab ini membahas mengenai penarikan data, pengumpulan data, ekplorasi data, dan

penyiapan data. Penarikan data yang dilakukan dengan menggunakan metode *web scraping* menggunakan selenium. Serta penyiapan data dengan melakukan proses *preprocessing* dan *postag*.

4. Bab 4 : Analisis Data dan Pengujian Model

Pada bab ini membahas mengenai analisa data dengan mencari kata kunci yang berada pada data sentimen positif dan negatif, dan membuat visualisasi *wordcloud* berdasarkan kata kunci tersebut. Setelah itu pembuatan model dilakukan dengan menggunakan model klasifikasi yaitu menggunakan *decision tree*, *logistic regression*, dan *random forest*, dan membahas mengenai hasil evaluasi model klasifikasi.

5. Bab 5 : Perancangan *Website* dan Implementasinya

Pada bab ini membahas mengenai perancangan perangkat lunak dengan membahas fitur-fitur yang dibutuhkan oleh pengguna dan implementasinya kepada perangkat lunak berdasarkan fitur-fitur tersebut.

6. Bab 6 : Kesimpulan dan Saran

Pada bab ini membahas mengenai kesimpulan yang telah dilakukan selama penelitian dan saran untuk para peneliti yang akan melakukan penelitian serupa.