Jurnal Mirai Manajemen Vol 7, No 1 (2022), *Pages **1 - 17***

ISSN : **2597 - 4084** Published By **STIE Amkop Makassar**

Analisis Sentimen *IMBd Film Review* Dataset Menggunakan *Support Vector Machine* (SVM) dan Seleksi *Feature Importance*

Hilda Nuraliza[1*], Oktariani Nurul Pratiwi[2], Faqih Hamami[3]

[1] Sistem Informasi, Fakultas Rekayasa Industri, Universitas Telkom

[2] Sistem Informasi, Fakultas Rekayasa Industri, Universitas Telkom

[3] Sistem Informasi, Fakultas Rekayasa Industri, Universitas Telkom

ABSTRACT

Perkembangan teknologi internet khususnya dalam bidang perfilman memberikan sarana terbuka bagi masyarakat dalam mengekspresikan pendapat dan emosional. Salah satu pendapat yang seringkali masyarkat keluarkan yaitu berupa penilaian sebuah film pada platform tertentu seperti IMDB. Ulasan yang dikeluarkan tentunya mengandung emosional yang dibawakan oleh masyarakat itu sendiri, baik emosional positif maupun negatif yang dinamakan sentimen. Sentimen atau opini masyarakat ini perlu dianalisis untuk mengklasifikasikan opini sesuai dengan kelasnya sehingga kecenderungannya terhadap suatu objek dapat diketahui. Adapun metode yang digunakan dalam penelitian ini yaitu metode *data mining* dengan Knowledge Discovery in Database (KDD). Tujuan dari penelitian ini yaitu analisa sentiment IMDB film review oleh masyarakat menggunakan algoritma *Support Vector Machine* dan seleksi *Feature importance* untuk memperoleh pola dan hasil yang terbaik. Dengan pengujian validasi akurasi data menggunakan metode split data sederhana dan *k-fold cross validation* yang menghasilkan akurasi sebesar 91.942% dan 87.699%. Lalu Kemudian dilakukan evaluasi menggunakan *confusion matrix* dengan penetapan max *feature* sebesar 10000 untuk memeriksa nilai ketepatan prediksi yang dilakukan oleh model yaitu diperoleh akurasi sebesar 88.033%. Dalam hal ini dapat dibuktikan bahwa kemampuan model dalam melakukan klasifikasi dinilai cukup baik.

Keywords:
Data Mining, KDD, *Feature Importance,* SVM, *Confusion Matrix*

1. Introduction

Film merupakan sebuah bentuk komunikasi massa elektonik yang berupa media pertunjukkan audio visual yang mampu menampilkan kata-kata, bunyi, citra, dan kombinasinya (Sobur, 2004). film mempunyai suatu dampak bagi penonton, dampak-dampak tersebut dapat berbagai macam seperti dampak psikologis dan dampak sosial (Effendy, 2003). Salah satu contoh dampak dari psikologis yang diterima oleh masyarakat yaitu berupa sudut pandang dan perspektif terhadap penilaian sebuah film yang diberikan oleh masyarakat berupa ulasan. Ulasan yang diberikan pada sebuah film tentunya dapat berupa ulasan positif dan ulasan *negative*. Didukung dengan perkembangan internet yang sangat pesat menjadi media sarana untuk masyarakat dalam menyalurkan pendapat dan penilaian terhadap film itu sendiri. Platform informatif yang berkaitan dengan film yang cukup popular adalah *Internet Movie Database* (IMDB). Platform ini merupakan sebuah situs website basis data informasi yang berkaitan dengan film, acara televise, video rumahan, dan permainan video.

Berbicara tentang data, data merupakan objek dengan dimensi yang sangat besar, baik dari banyaknya instan maupun atribut yang dimilikinya. Misalnya, untuk data *review film*, satu *film* bisa memiliki banyak atribut. *Film* dapat memiliki data judul, *genre*, penulis, sutradara, produser, dan lain-lain. Banyaknya atribut yang bisa dimiliki oleh satu objek belum tentu merupakan informasi relevan yang dibutuhkan oleh sistem *data mining*. Untuk itulah perlu dilakukan proses reduksi data. *Feature importance* merupakan cara yang efektif untuk melakukan reduksi data dan menjadi langkah penting yang perlu dilakukan supaya aplikasi *data mining* berhasil dengan baik.

Feature importance merupakan proses pemilihan subset dari fitur/atribut yang optimal dengan menggunakan kriteria tertentu. *Feature importance* merupakan salah satu dari proses *pre-processing* pada suatu dataset yang akan dilakukan proses *data mining*. Dengan melakukan *feature importance* ini mampu untuk mengurangi jumlah *feature* yang tidak relevan, menghilangkan redundansi data, menghilangkan *feature* yang mengandung *noisy* dan akan memberikan efek meningkatkan kecepatan dalam melakukan *data mining*, meningkatkan akurasi *learning*, dan menghasilkan model yang baik (Liu, Huan dkk. 2010). Hal ini, sangat mendukung proses *data mining* pada dataset dengan jumlah yang besar seperti IMDB *movie review dataset*. pada penelitian ini akan dilakukan proses *text mining* salah satunya menggunakan algoritma SVM dan pemilihan fitur pada analisis sentiment IMDB dataset.

Sentimen Analysis yaitu suatu penelitian dalam bidang keilmuan *Machine Learning* yang membahas tentang opini dalam bentuk teks. Dengan menggunakan sebuah pendekatan yang mendefinisikan bahasa tersebut ke arah positif dan negatif. (Sing,V.K., Piryani,R., Uddin, A., Waila,

P., 2013). Dalam pengambilan data opini tersebut dapat menggunakan proses scrapping data yang selanjutnya dilakukan proses labeling untuk memperoleh opini *positive* dan *negative*. Algoritma yang dapat digunakan pada kasus ini salah satunya adalah algoritma *Support Vector Machine* (SVM).

Algoritma SVM menghasilkan nilai akurasi tinggi seperti menguji sentimen terhadap wacana politik pada media sosial *online*, analisis sebuah sentimen komentar mahasiswa pada sistem pembelajaran di perguruan tinggi, analisis pada *tweets* di *Twitter* yang mengeluarkan opini tentang produk mobil otomatis dan produk Apple, dan rata-rata tingkat akurasi yang didapatkan adalah 50 %-90 % (Hidayat A.N, 2015).

Berdasarkan informasi yang telah dijelaskan, penulis mencoba menggunakan Algoritma SVM dan *feature importance* untuk pengolahan data dengan topik permasalahan yang akan diangkat yaitu mengenai analisis sentimen IMDB dataset dengan objek data *movie reviews*. Hasil ulasan film yang diberikan oleh masyarakat terhadap suatu film ini sangat menarik untuk dianalisis terkait berbagai opini dan fenomena. Penelitian ini bermaksud untuk mengenalisis kecenderungan penilaian terhadap suatu topik dalam hal ini adalah mengenai film pada platform IMDB. Proses *text mining* ini tentunya menggunakan metode NLP (*Natural Process Language*).

2. Literature Review

2.1. Data Mining

Data mining merupakan sebuah proses dalam mengekstraksi sebuah data yang sebelumnya bersifat implisit, tidak terstruktur menjadi sebuah informasi dan pengetahuan atau pola data terstruktur yang menjadi lebih berguna (Witen, Ian H. Frank, 2011). *Data mining* sendiri merupakan bagian dari *Knowledge discovery in databases (KDD*). *Knowledge Discovery* merupakan pencarian informasi atau pola tertentu yang tersembunyi dalam suatu data. Ilustrasi dari proses *data mining* dapat dilihat dari gambar 2.1 dibawah ini. Dimana gambar tersebut menunjukkan bahwa data-data yang bersifat mentah dan dianggap sampah karena tidak terstruktur dan terpola akan diolah hingga membentuk sebuah informasi dan pengetahuan atau pola baru yang menjadi lebih berguna.

Gambar II.1 Ilustrasi Data mining

Sumber : Buku *Data mining: Algoritma dan Implementasi dengan Pemrograman PHP*

2.2. Metode Classification

Classification adalah Sebuah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data dengan tujuan untuk memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Hal ini juga dapat dikatakan sebagai pembelajaran (klasifikasi) yang memetakan sebuah unsur (item) data kedalam salah satu dari beberapa kelas yang sudah didefinisikan (Bustami, 2013). Klasifikasi banyak digunakan dalam berbagai aplikasi, diantaranya adalah untuk mendeteksi kecurangan (fraud detection), pengelolaan pelanggan, diagnosis medis, prediksi penjualan dan lain sebagainya.

Klasifikasi data terdiri dari 2 langkah. Langkah pertama yaitu melakukan pelatihan pada dataset yang sudah ada. Data latih (*training set*) yang digunakan sudah memiliki label kelas. Inilah yang membedakan klasifikasi dengan clustering, dimana klasifikasi membutuhkan proses pelatihan data sehingga memerlukan data latih yang sudah mengandung label kelas. Sedangkan pada *clustering*, tidak ada label kelas. Karena pada sudah tersedia label kelas, klasifikasi disebut sebagai supervised learning.

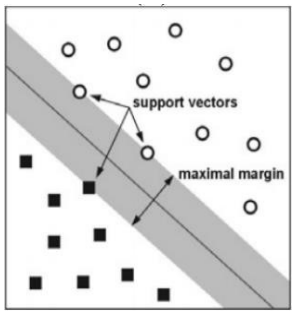
2.3. Algoritma SVM

SVM merupakan teknik supervised machine learning yang dikembangkan oleh Vapnik tahun 1995 dan dikembangkan lebih lanjut oleh Joachims tahun 1998. Berbeda dengan metode terdahulu SVM memiliki konsep dan teori yang terstruktur dan baik sehingga metode ini adalah metode dengan akurasi terbaik dalam bidang klasifikasi *text* yang digunakan dalam berbagai macam kasus untuk sentiment klasifikasi. SVM bekerja dengan membagi data *train*ing menjadi 2 kelas dengan memperkirakan garis hyperplane dan mencari jarak maksimal dari *hyperplane* ke *data training* terdekat agar didapatkan generalisasi untuk proses klasifikasi dengan data *test*. Berdasarkan Gambar 2.3 garis *hyperplane* berada pada tengah dari nilai maksimal margin dari data terdekat. Persamaan untuk mendapatkan *hyperline*:

$$\omega \cdot \chi = b = 0$$

Formula menghitung maksimal *margin* adalah:

$\omega \cdot \chi - b = 1$, dan $\omega \cdot \chi - b = -1$



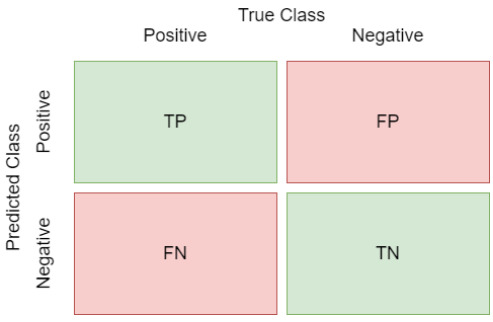
Gambar II.2 Ilustrasi SVM

2.4. Feature Importance

Feature importance (pemilihan fitur) adalah salah satu teknik *Data Preparation* terutama dalam lingkup data reduction. Hal ini karena *feature importance* melakukan pemilihan atribut-atribut yang paling mempengaruhi hasil dari klasifikasi. Dengan sistem pemilihan ini maka akan menyebabkan atribut-atribut yang ada dalam dataset yang ada berkurang sehingga bisa disebut kita melakukan reduksi data. Pada proses pemilihan atribut ini dilakukan dengan tetap mempertahankan informasi-informasi yang penting. Sehingga tidak akan banyak mempengaruhi hasil klasifikasi nantinya. Dengan melakukan pemilihan fitur ini, hasil yang diharapkan setelah melakukan klasifikasi, performansi yang dihasilkan menjadi lebih baik. Popularitas data besar baru-baru ini menghadirkan tantangan unik untuk pemilihan fitur tradisional (Li dan Liu 2017), dan beberapa karakteristik data besar seperti kecepatan dan variasi memerlukan pengembangan algoritma seleksi fitur baru.

2.4. Confusion Matrix

Penghitungan metrik dilakukan dengan dasar awal *confusion matrix* seperti pada gambar di bawah.



Gambar II.3 Ilustrasi Predictive Class

TP: *True Positive* (jumlah prediksi benar kelas positif) TN: *True Negative* (jumlah prediksi benar kelas negatif)

FN: *False Negative* (kelas asli positif diprediksi negatif) FP: *False Positive* (Kelas asli *negative*, diprediksi positif)

- *Accuracy*

Metrik yang menunjukkan berapa banyak kelas yang diprediksi dengan benar oleh model. Merupakan model paling sering dipakai namun pada kasus data tidak seimbang memiliki kerentanan terhadap salahnya interpretasi performa (bias). Persamaan dari metrik ini sendiri adalah sebagai berikut:

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

- *Precision*

Metrik yang menunjukkan seberapa benar hasil prediksi kelas positif model sebagai kelas positif (aktual). Namun metrik ini tidak dapat menggambarkan secara jelas kelas positif (aktual) memiliki berapa banyak hasil prediksi benar. Persamaan dari metrik ini sendiri adalah sebagai berikut:

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

- *Recall (True Positive Rate (TPR))*

Metrik yang menunjukkan seberapa benar kelas positif (aktual) diprediksi sebagai kelas positif. Namun metrik ini tidak dapat menggambarkan seberapa baik model melakukan prediksi kelas positif sebagai kelas positif (aktual). Persamaan dari metrik ini sendiri adalah sebagai berikut:

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

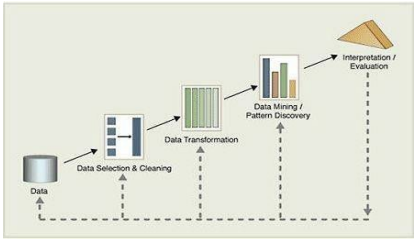
- *F1 Score*

Metrik yang menutupi kekurangan pada *precision* dan *recall* di dalam penilaian performa terhadap kelas positif dengan cara menghitung rata-rata *harmonic* dari keduanya. Namun dikarenakan kedua metrik sebelumnya hanya berfokus pada kelas positif menyebabkan f1 *score*juga tidak dapat menggambarkan secara spesifik penilaian performa terhadap kelas negatif. Akan tetapi semua hal tersebut diatasi dengan menerapkan versi weighted yang memperhitungkan keseluruhan kelas yang ada beserta distribusinya. Persamaan dari metrik ini sendiri adalah sebagai berikut:

$$f1 - score = 2 * \frac{precision * recall}{precision + recall} = 2 * \frac{2TP}{2TP + FP + FN}$$

3. Method, Data, and Analysis

Dalam proses penelitian, penulis menggunakan *Knowledge Discovery In Database* (KDD) sebagai kerangka pemecahan masalah. Metode KDD merupakan keseluruhan proses nontrivial untuk mencari dan mengidentifikasi pola (*pattern*) dalam data, dimana pola yang ditemukan bersifat sah, baru, dapat bermanfaat dan dapat dimengerti. KDD berhubungan dengan teknik integrasi dan penemuan ilmiah, interprestasi dan visualisasi dari pola-pola sejumlah kumpulan data. (Fayyad, 1996). Penulis menggunakan metode *Knowledge Discovery In Database* (KDD) dengan tujuan untuk memeriksa basis data berukuran cukup besar sebagai cara untuk menemukan pola atau *pattern* yang baru dan berguna.



Gambar III.1 Data mining Stages Process

Source : Data Mining Books: Algorithms and Implementation with PHP Programming

Pada penelitian ini mengacu pada gambar diatas, dalam memetakan tujuan terhadap objek penelitian berikut adalah tahapan proses *data mining* yang dilakukan oleh peneliti.

1. Tahapan pertama yaitu proses pengumpulan data atau dataunderstanding.
2. Tahapan kedua yaitu *preprocessing data*, pada proses preparasi data ini melakukan proses persiapan data sebelum dilakukan proses pelatihan model diantaranya: *Data selection* dan *data cleaning* perlu dilakukan untuk memperoleh data yang terstruktur dan bersih sehingga terhindar dari redundansi, anotasi, *missing valu*e, dan kecacatan data lainnya sehingga dalam proses *data mining* tidak terjadi *error* dan kecacatan yang fatal.
3. Tahapan ketiga yaitu *transformation data*, Proses ini merupakan proses Proses konversi atau encoding adalah proses mengubah data menjadi format tertentu agar dapat digunakan dan dilacak nantinya.
4. Tahapan keempat yaitu pemrosesan data mining dengan output pattern. Pada proses ini, pemrosesan *data mining* dilakukan menggunakan algoritma *Support Vector Machine* (SVM) yang didukung dengan proses *feature importance* untuk mencari pola. Pola serta data yang dicari merupakan pola dan data yang akan menarik.
5. Tahapan kelima yaitu evaluasi dan interpretasi. Setelah pola dan data ditemukan, selanjutnya adalah dengan menampilkan data tersebut dalam bentuk yang mudah dipahami oleh semua pembaca, pengguna atau pihak yang memiliki kepentingan. Hal ini dapat disebut dengan visualisasi data, visualisasi data dapat dipersembahkan dalam bentuk diagram yang tentunya

berguna bagi pihak yang berkepentingan. Hasil evaluasi ditampilkan ke dalam metrik evaluasi dengan metode proses *feature importance.*

4. Result and Discussion

Berdasarkan metode yang telah dirumuskan, berikut adalah tahapan dalam penelitian.

1. Pengumpulan Data

   Proses pengumpulan data terdiri dari 2 (dua) tahap, yaitu pemahaman data dan penjelasan data.

   - Data Understanding

     Data yang dipilih adalah data sekunder yang artinya data tersebut telah dikumpulkan oleh pihak kedua yaitu berupa dataset Movie Review yang diperoleh dari website IMDB. Data ini merupakan data review yang diberikan oleh masyarakat atau penonton pada judul film tertentu. Dataset ini bersumber dari data opensource Kaggle yang dikembangkan pada tahun 2011 (ACL,2011). Dataset adalah file dengan format . CSV yang terdiri dari total 50.000 baris data dan 2 kolom atribut setelah proses penggabungan seluruh dataset review setiap film dengan label yang terdapat pada sentimen yang menunjukkan hasil review baik atau burk sebuah film di IMDB yang dapat dilihat pada tabel dibawah ini.

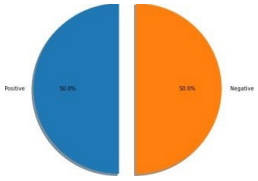
     Tabel 1 tabel deskripsi data

No	Atribut	Jumlah Baris
1	Review	50.000
2	Sentiment	50.000

     Data tersebut merupakan data yang siap untuk diolah dan diklasifikasikan berdasarkan sentimen positif dan sentimen negatif, sehingga dapat dilakukan proses pengolahan selanjutnya yaitu proses data mining.

   - *Exploratory Data Analysis*

     Hasil eksplorasi menunjukkan sebaran fitur label kelas positif dan negatif. Terlihat bahwa sebaran pada fitur label kelas positif dan negatif menunjukkan data keseimbangan dimana kelas positif 50% dan kelas negatif 50%.

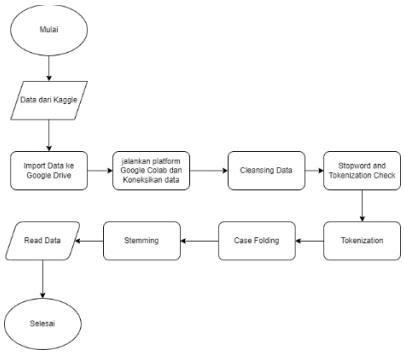


Gambar 2 Distribusi label

dan analisis terhadap penemuan-penemuan penelitian, penjelasan serta penafsiran dari data dan hubungan yang diperoleh, serta pembuatan generalisasi dari penemuan. Apabila terdapat hipotesis, maka pada bagian ini juga menjelaskan proses pengujian hipotesis beserta hasilnya.

2. Preprocessing Data

Pada tahap ini penulis melakukan pembersihan dan penyempurnaan data untuk memastikan apakah data yang akan digunakan berkualitas baik. Dalam proses ini, ada enam hal yang perlu dipastikan, yaitu keakuratan data, kelengkapan, konsistensi, ketepatan waktu, keandalan, dan dapat diinterpretasikan dengan baik. Dalam penelitian ini, penulis menggunakan platform google colab sebagai media preprocessing data. Adapun beberapa tahapan preprocessing data yang dilakukan oleh penulis dapat diperlihatkan pada diagram alir berikut ini.



Gambar 3 Flowchart Prapemrosesan Data

Tabel berikut adalah proses dari tahapan preprocessing yang menunjukkan perbedaan dari hasil setelah preprocessing dilakukan.

Tabel 2 Tahapan Data Preprocessing

Fase	Exist	Output

Cleansing Data		['one', 'reviewers', 'mentioned','watching', 'just', 'oz', 'go', 'violence', 'its', 'hardcore','classic', 'use'], etc
Stopword and Tokenization Check	['I', 'seen', 'film','times', 'I', 'excited', 'acting','perfect', 'romance', 'joe', 'jean', 'keeps', 'edge ', 'seat', 'plus', 'I', 'think', 'bryan', 'brown', 'tops', 'brilliant','film']	'seen', 'film', 'times', excited','acting', 'perfect', 'romance', 'keeps', 'edge', 'seat', 'plus', 'think', 'tops', 'brilliant','film'
Tokenization	i seen film times i excited acting perfect romance joe jean keepsedge seat plus ithink bryan brown tops brilliant film	['I', 'seen', 'film', 'times', 'I', 'excited', 'acting','perfect', 'romance', 'joe', 'jean', 'keeps', 'edge ', 'seat','plus', 'I', 'think', 'bryan', 'brown', 'tops', 'brilliant','film']
Case Folding	I Seen Film Times I Excited Acting Perfect Romance Joe Jean Keeps	i seen film times i excited actingperfect romance joe jean keeps

3. Pembobotan TF-IDF

   Setelah proses *preprocessing data* selesai dilakukan, langkah selanjutnya yang dilakukan pada penelitian ini yaitu proses TF-IDF Victorizer. Pada penelitian ini, TF digunakan untuk menghitung jumlah kemunculan kata dalam satu data dan IDF digunakan untuk menghitung kata yang muncul lebih dari satu data ulasan yang dianggap umum dan dinilai tidak penting. Berikut ini merupakan tahapan dari process TF-IDF yang dilakukan pada penelitian dengan mengambil beberapa sampel dokumen.

D1 = "*if like original gut wrenching laughter like movie if young old love movie hell mom liked br br great camp*"

D2 = "*it terrific funny movie does make smile what pity this film boring long it simply painfull the story staggering goal br br you feel better finish*"

D3 = "*a rating does begin express dull depressing relentlessly bad movie*"

Dari ketiga dokumen tersebut, maka akan diekstrak nilai TF-IDF yang telah disajikan pada tabel dibawah ini.

Tabel 3 pembobotan TF-IDF

Term	TF			df	D/Df	IDF	Wdt = TFdt * IDFft		
	D1	D2	D3				D1	D2	D3
if	2	0	0	2	1,5	0,176091	0,352183	0	0
like	2	0	0	2	1,5	0,176091	0,352183	0	0
original	1	0	0	1	3	0,477121	0,477121	0	0
gut	1	0	0	1	3	0,477121	0,477121	0	0
wrenching	1	0	0	1	3	0,477121	0,477121	0	0
laughter	1	0	0	1	3	0,477121	0,477121	0	0
movie	2	1	1	4	0,75	-0,12494	-0,24988	-0,12494	-0,12494
young	1	0	0	1	3	0,477121	0,477121	0	0
old	1	0	0	1	3	0,477121	0,477121	0	0
love	1	0	0	1	3	0,477121	0,477121	0	0
hell	1	0	0	1	3	0,477121	0,477121	0	0
mom	1	0	0	1	3	0,477121	0,477121	0	0
liked	1	0	0	1	3	0,477121	0,477121	0	0
great	1	0	0	1	3	0,477121	0,477121	0	0
camp	1	0	0	1	3	0,477121	0,477121	0	0
terrific	0	1	0	1	3	0,477121	0	0,477121	0
funny	0	1	0	1	3	0,477121	0	0,477121	0
does	0	1	1	2	1,5	0,176091	0	0,176091	0,176091
make	0	1	0	1	3	0,477121	0	0,477121	0
smile	0	1	0	1	3	0,477121	0	0,477121	0
pity	0	1	0	1	3	0,477121	0	0,477121	0
this	0	1	0	1	3	0,477121	0	0,477121	0
film	0	1	0	1	3	0,477121	0	0,477121	0

4. Splitting Data
   - Split data sederhana

     Pada pengujian pertama, yaitu dengan melakukan splitting data menjadi data *train* dan data *test* dengan rasio perbandingan 50:50. Kedua data ini tentu sudah melalui proses *preprocessing data* dan pemilihan feture importance yang dilakukan menggunakan *tools* google colab dan melakukan pemodelan menggunakan bahasa pemrograman python untuk memperoleh nilai akurasi. Maka diperoleh *score* data *train* dan data *test* sebagai berikut.

Tabel 4 Hasil dari Skenario

Splitting Data Score	
Split Data	Accuracy
Train	0.95849
Test	0.88034
Rata-rata	0.91942

Pada tabel diatas menunjukkan bahwa nilai akurasi data *train* sebesar 0.95849 atau setara dengan 95.849 % dan nilai akurasi data *testing* sebesar 0.88034 atau setara dengan 88.034 %. Hingga diperoleh rata-rata dari keduanya adalah 0.91942 atau setara dengan 91.942%.

- K-fold Cross Validation

  Pada pengujian kedua, yaitu dengan melakukan splitting data menjadi data *train* dan data *test* dengan rasio perbandingan 50:50. Dan tentunya kedua data ini telah melalui proses *preprocessing data* dan pemilihan feture importance menggunakan *tools* google collab dan akan melakukan pemodelan menggunakan metode *K-fold cross validation* dengan jumlah iterasi sebanyak 4 (empat) kali. Maka diperoleh *score*data *test* yang dapat dilihat pada gambar berikut ini.

Tabel 5 Hasil pengujian kfold

Split	Nilai	Persentase
Fold 1	0.87874	87.874%
Fold 2	0.87542	87.542%
Fold 3	0.87942	87.942%
Fold 4	0.87440	87.440%
Rata-Rata	0.87699	87.699%

Pada tabel V.2 menunjukkan sampel dari akurasi, presisi, recall dan f1-*score*setelah dilakukan proses *modelling* menggunakan metode *K-fold cross validation*. Untuk hasil pengujian fold dapat dilihat pada tabel dibawah ini.

5. Pemodelan Data

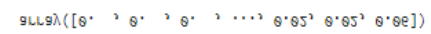
Data *test* yang telah dilakukan pembobotan menggunakan TF-IDF kemudian akan diklasifikasikan menggunakan algoritma *Support Vector Machine* (SVM). Pada tahap klasifikasi ini, data yang akan dihitung merupakan data kalimat dalam setiap dokumen. Data masukan secara bersamaan akan diubah menjadi data vector. Pada penelitian ini, format representasi data yang digunakan adalah format SVM light seperti : "1 1: 0,477121". Angka 1 pada karakter pertama menyatakan data tersebut masuk kedalam label positif sedangkan bila angka -1 pada karakter pertama menyatakan data tersebut masuk kedalam label negatif dan angka satu sebelum tanda ":" merupakan indeks dan angka setelah ":" merupakan bobot TF-IDF dari term tersebut. Berikut adalah contoh pengubahan data teks menjadi data vector yang disajikan kedalam tabel dibawah ini.

Tabel 4 Translasi ke Data Vektor

Dokumen	if like original gut wrenching laughter like movie if young old love movie hell mom liked br br great camp
Vektor	[-1 1:0,352183 2:0, 352183 3:0,477121 4:0,477121 5:0, 477121 6:0,477121 7:0, 352183 8:-0,24988 9:0,352183 10:0, 477121 11:0, 477121 12:0, 477121 13:-0,24988 14:0, 477121 15:0, 477121 16:0, 477121 17:0, 477121 18:0, 477121]

6. Proses Seleksi *Feature importance*

Proses selanjutnya pada penelitian ini adalah proses seleksi *feature importance*, tujuannya adalah untuk menentukan *feature* penting yang akan digunakan sehingga dapat melihat seberapa kuat variable *input* mempengaruhi hasil prediksi.



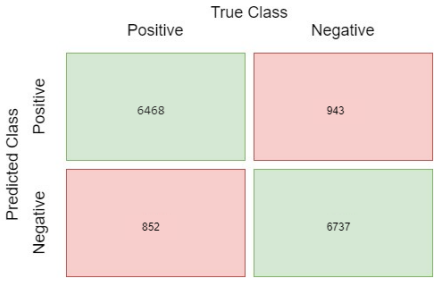
Gambar 3 Nilai *Feature importance*

Pada gambar 3 menunjukkan nilai dari *feature importance* yang telah di urutkan dan di representasikan kedalam variable "thresholds" yang digunakan dalam proses *train*ing untuk membangun model. Pada penelitian ini, *feature importance* digunakan untuk mengukur kepentingan variable independent terhadap ketepatan opini. Semakin besar nilai *feature importance* menunjukkan semakin besar juga veran variable tersebut dalam menganalisis. Nilai "thresholds" diatas di peroleh dari perhitungan mean decrease in impurity (MDI).

7. Evaluasi

Setelah dilakukan validasi data untuk memperoleh nilai *score* yang dihasilkan melalui proses seleksi *feature importance* dan berdasarkan metode yang dipilih yaitu split data sederhana dengan *K-fold cross validation*. Sebagai proses terakhir dari penelitian ini adalah melakukan

evaluasi model untuk memvalidasi kesesuaian akurasi, presisi, recall, dan f1-*Score* yang dihasilkan oleh model. Maka diperoleh *score* akurasi yang dihasilkan oleh terminal seperti dibawah ini.



Gambar 3 *Confusion matrix*

Pada gambar V.3 dapat dilihat bahwa kemampuan model dalam melakukan klasifikasi dengan prediksi benar positif adalah sebesar 6468 dan benar negatif adalah 6737. Sedangkan kemampuan model dalam memprediksi salah positif adalah sebesar 943 dan salah negatif adalah 852. Artinya model tersebut melakukan klasifikasi dan memberikan ketepatan prediksi yang cukup baik. Untuk mengetahui nilai akurasi yang dihasilkan dari *confusion matrix* maka dapat dilihat pada classification report dengan perbandingan 50:50 dapat dilihat pada gambar dibawah ini.

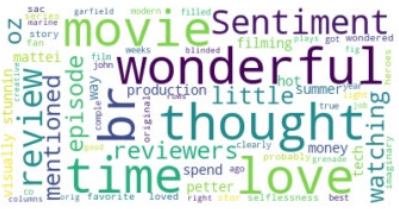
```
Classification report :
0.8803333333333333
              precision    recall  f1-score   support

           0       0.88      0.87      0.88      7411
           1       0.88      0.89      0.88      7589

    accuracy                           0.88     15000
   macro avg       0.88      0.88      0.88     15000
weighted avg       0.88      0.88      0.88     15000
```

Gambar 4 Grafik classification report

Pada gambar V.4 menunjukkan bahwa nilai akurasi, precision, recall, dan F1-*Score*yang dihasilkan dari model yaitu sebesar 0,8803 atau setara dengan 88.03%. Maka dari itu, kinerja yang dilakukan oleh model klasifikasi dengan pemilihan *feature importance* terbukti memberikan hasil akurasi yang cukup baik.

8. Visualisasi Sentimen

   Setelah dilakukan pengujian untuk memperoleh nilai akurasi dari masing-masing model, selanjutnya adalah visualisasi sentiment menggunakan *word cloud*.



Gambar 5 *word cloud* sentimen positif

Pada gambar V.5 menunjukkan bahwa kata ulasan positif penonton terhadap sebuah film yang seringkali muncul dan terangkum pada website IMDB adalah : *"Woderful"* (film tersebut sangat indah dan bagus); *"Love"* (penonton menyukai film tersebut); *"Visually"* (secara visual, film tersebut sangat bagus); *"Favorite"* (film tersebut menjadi favorit penonton); *"Best"* (penonton sepakat bahwa film tersebut memang terbaik); *"Wondered"* (Film tersebut sukse membuat penonton bertanya-tanya dan sulit menebak *ending* film tersebut).

Gambar 6 *word cloud* sentimen negatif

Pada gambar 6 menunjukkan bahwa kata-kata ulasan *negative* penonton terhadap sebuah film yang sering muncul dan terangkum pada website IMDB adalah : *"Bad"* (film tersebut buruk dan tidak bagus untuk ditonton) ; *"Scariest"* (film tersebut paling seram dan tidak baik untuk dilihat anak-anak) ; *"Plot"* (penonton depakat jika alur film tersebut tidak jelas arahnya) ; *"Diasgeree"* (beberapa penggalan cerita film tersebut banyak yang tidak disetujui oleh penonton) ; *"Expect"* (Ending dan jalan cerita film tersebut tidak sesuai dengan ekspektasi penonton).

5. Conclusion and Suggestion

Berdasarkan hasil penelitian tugas akhir yang telah dilakukan oleh penulis, maka diperoleh rumusan simpulan sebagai berikut.

1. Pada penelitian ini menunjukkan bahwa metode seleksi *feature importance* mampu memberikan hasil klasifikasi dan prediksi sentiment analisis pada IMDB movie review. Dengan perbandingan data *train* dan *test* sebesar 50:50 yang disajikan kedalam *confusion matrix* membuktikan bahwa model melakukan klasifikasi dengan baik.
2. Pola kata yang dihasilkan berdasarkan hasil proses klasifikasi diperoleh sentiment positif dan negatif yang direpresentasikan kedalam bentuk *word cloud*. Keduanya membentuk kumpulan

kata yang berkaitan dengan kategori sentiment *positive* dan negatif yang menunjukkan frekuensi kata yang sering di sebutkan dalam ulasan oleh penonton.

Reference :

Sentiment Analisis Review Film Di IMDB Menggunakan Algoritma SVM2019JURNAL SISTEM INFORMASI DAN TEKNOLOGI INFORMASI47-56

2011Analisis dan Implementasi Algoritma ReliefF untuk Feature importance pada Klasifikasi Dataset MulticlassBandungZ.K. Abdurahman Baizal, Erda Guslinar Perdana

ANALISIS PERBANDINGAN ALGORITMA NAIVE BAYES DAN SUPPORT VECTOR MACHINE DALAM MENGKLASIFIKASIKAN JUMLAH PEMBACA ARTIKEL

ONLINE 2018Jurnal Teknik Informatika (JIKA) Universitas Muhammadiyah Tangerang62 - 72

Hidayat, A. N. (2015). Analisis Sentimen Terhadap Wacana Politik Pada Media Masa Online Menggunakan Algoritma Support Vector Machine Dan Naive Bayes. Jurnal Elektronik Sistem Informasi dan Komputer, 1(1), 12-18.

Habibi, M. (2017). Analisis Sentimen dan Klasifikasi Komentar Mahasiswa pada Sistem Evaluasi Pembelajaran Menggunakan Kombinasi KNN Berbasis Cosine Similarity dan Supervised Model (Doctoral dissertation, Universitas Gadjah Mada).

Saleh, M. R., Martín-Valdivia, M. T., Montejo-Ráez, A., & Ureña-López, L. A. (2011). Experiments with SVM to classify opinions in different domains. Expert Systems with Applications, 38(12), 14799-14804.

Bustami, B. (2013). Penerapan algoritma Naive Bayes untuk mengklasifikasi data nasabah asuransi. TECHSI-Jurnal Teknik Informatika, 5(2).

Danang Aji Irawan[1], Z. A. (2011). ANALISIS DAN IMPLEMENTASI ALGORITMA RELIEF UNTUK FEATURE IMPORTANCE PADA KLASIFIKASI DATASET

MULTICLASS. Bandung: Open Library Telkom University.

Liu, H., Motoda, H., Setiono, R., & Zhao, Z. (2010, May). Feature selection: An ever evolving frontier in data mining. In Feature selection in data mining (pp. 4-13). PMLR.

Singh, V. K., Piryani, R., Uddin, A., & Waila, P. (2013, March). Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In 2013 International mutli-conference on automation, computing, communication, control and compressed sensing (imac4s) (pp. 712-717). IEEE.

Chandani, V., Wahono, R. S., & Purwanto, P. (2015). Komparasi algoritma klasifikasi Machine Learning dan feature selection pada analisis sentimen review film. Journal of Intelligent Systems, 1(1), 56-60.

Rahutomo, F., Saputra, P. Y., & Fidyawan, M. A. (2018). Implementasi Twitter Sentiment Analysis Untuk Review Film Menggunakan Algoritma Support Vector Machine. Jurnal Informatika Polinema, 4(2), 93-93.

Kharde, V., & Sonawane, P. (2016). Sentiment analysis of twitter data: a survey of techniques. arXiv preprint arXiv:1601.06971.

Putranto, H. A., Setyawati, O., & Wijono, W. (2016). Pengaruh Phrase Detection dengan POS-Tagger terhadap Akurasi Klasifikasi Sentimen menggunakan SVM. Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI), 5(4), 252-259.

Suntoro, J. (2019). Data mining: Algoritma dan Implementasi dengan Pemrograman PHP. Jakarta: PT Elex Media Komputindo.

Chandani, V., Wahono, R. S., & Purwanto, P. (2015). Komparasi algoritma klasifikasi Machine Learning dan feature selection pada analisis sentimen review film. Journal of Intelligent Systems, 1(1), 56-60.

Elfaladonna, F., & Rahmadani, A. (2019). Analisa Metode Classification-Decission Tree dan Algoritma C. 45 untuk Memprediksi Penyakit Diabetes dengan Menggunakan Aplikasi Rapid Miner. SINTECH (Science And Information Technology) Journal, 2(1), 10-17.

Ikhsan Subagyo, L. D. (2019). Sentiment Analisis Review Film Di IMDB Menggunakan Algoritma SVM. JURNAL SISTEM INFORMASI DAN TEKNOLOGI INFORMASI, 47 - 56.