**ANALISIS SENTIMEN ULASAN FILM IMDB MENGGUNAKAN TF-IDF DAN SUPPORT VECTOR MACHINES** *]*

**Auliyah Afifah Adnan Hakim1, Fiqrah Idhul Dwi Anugerah2, La Ode Muhammad Yudhy Prayitno3**

1,2 Teknik Informatika, Teknik, Universitas Halu Oleo.

1alamat e-mail, 2 emailnyafiqrah10@gmail.com, 3 mr.yudhyt@gmail.com

**Abstrak**

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.Banyak peneliti saat ini menggunakan metode pembelajaran mesin yang diawasi untuk mempelajari analisis sentimen. Analisis dapat dilakukan pada ulasan film, Support Vector Machines (SVM) digunakan untuk menganalisis kumpulan data dengan menggunakan parameter yang berbeda. Analisis dan diskusi dilakukan untuk memungkinkan kesimpulan bahwa SVM telah berhasil diimplementasikan dengan menggunakan data IMDb untuk penelitian ini, 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.

**Kata kunci** : film, SVM, text-mining, TF-IDF

1. **Pendahuluan**

Film adalah jenis komunikasi massa elektonik yang berupa media pertunjukan audio visual yang dapat menampilkan kata-kata, bunyi, gambar, dan kombinasi dari semua ini. Film memiliki dampak pada penonton, yang dapat beragam. seperti dampak sosial dan psikologis. Salah satu contoh dampak psikologis yang diterima oleh masyarakat adalah sudut pandang dan perspektif yang dimiliki masyarakat tentang bagaimana mereka menilai sebuah film; ulasan yang diberikan tentunya dapat berupa ulasan positif atau negative. Dibantu oleh perkembangan internet yang sangat pesat, ini menjadi media yang memungkinkan masyarakat untuk berbagi pendapat dan penilaian tentang film itu sendiri. Internet Movie Database (IMDB) adalah situs web yang menawarkan informasi tentang film-film populer(Nur Cahyo et al., 2023).

Ada banyak penelitian dan kemajuan teknologi yang mendorong kebutuhan akan informasi. Saat ini, informasi yang kita miliki sangat beragam dan luas sehingga diperlukan berbagai metode penyampaian informasi untuk mencakup keseluruhannya.   
Semua pengguna internet yang aktif memiliki hak untuk menyatakan pendapat mereka tentang masalah apa pun. Resensi film adalah salah satu jenis resensi. Banyak dokumen teks berasal dari berbagai sumber, termasuk informasi, review, pendapat, artikel, buku, perpustakaan digital, pesan email, dan halaman web. Karena kemudahan penggunaan web, semakin

banyak dokumen teks yang berisi pendapat dan opini. Text mining adalah teknik penggalian teks terbaru. Sangat penting untuk memberikan dokumen teks dari website yang berisi pendapat, komentar, kritik, dan review, karena jika dikelola dengan baik, dapat memberikan informasi yang bermanfaat untuk membantu orang atau organisasi membuat keputusan. Text mining adalah proses pengambilan intisari dari dokumen teks untuk menghasilkan hasil yang bermanfaat. Ada beberapa algoritma atau metode untuk analisis sentimen, antara lain Support Vector Machine (SVM) dan term frequency–inverse document frequency.

Di dalam penelitian ini algoritma yang digunakan adalah TF-IDF sebagai pembobotan kata dan Support Vector Machine (SVM) untuk pembagian data.

1. **Metode Penelitian**
   1. **Text Mining**

Text mining merupakan proses ekstraksi berita dari data asal yang belum terstruktur. Data yang belum terstruktur akan diolah memakai teknik serta metode tertentu membuat berita yang bermanfaat untuk pengguna. Text mining adalah teknik yang dipergunakan untuk menangani persoalan classification, clustering, information extraction, serta information retrieval

* 1. **Analisis Sentimen**

Analisis sentimen adalah teknik mengetahui, mengekstrak, dan menjalankan informasi tekstual secara otomatis untuk menemui keterangan sentimental dalam mengungkapkan pemikiran. Sentiment analysis dipergunakan untuk melihat pendapat atau kesamaan opini terhadap sebuah persoalan atau objek oleh seorang menuju ke opini positif atau negatif

* 1. **Prepocessing**

Preprocessing adalah proses normalisasi term dari sebuah kalimat. Hal ini dilakukan untuk memperoleh data pelatihan yang baik dan menyinkronkan fitur-fitur yang diekstraksi dengan yang diinginkan sehingga pengolahan data dapat disederhanakan. Mengumpulkan data opini dari media sosial di Twitter mungkin tidak sama dengan kata baku, kata-kata yang tidak ada dalam kamus, atau bahasa daerah yang digunakan atau dihilangkan. Diperlukan preprocessing atau normalisasi untuk mengatasi masalah ini, guna menghilangkan ekspresi atipikal dan mengembalikan bagian teks ke teks natural guna meminimalisir noise pada tahap selanjutnya. Sesuai contoh pada buku Pra-Pemrosesan, terdapat enam proses yang terdiri dari(Kardila, 2023):

1. *Normalization*

Tahap ini dilakukan agar kata-kata yang disingkat atau diperpanjang menjadi kata-kata yang normal sesuai dengan Kamus Besar Bahasa Indonesia (KBBI). Konversi slang word adalah proses mengubah kata tidak baku menjadi kata baku.

1. *Cleaning*

Pada tahap ini, karakter non-abjad dihilangkan untuk mengurangi noise. Karakter yang dihilangkan antara lain tanda baca seperti titik (.), koma (,), tanda tanya (?), dan tanda seru (!), simbol seperti tanda “@” pada nama pengguna, dan hashtag (#), Emoji, Situs Web – Alamat

1. *Case Folding*

Case folding adalah tahap untuk mengkonversi karakter alfabet yang telah melalui tahap cleansing ke huruf kecil (lower case).

1. *Tokenizing*

Tahap ini berfungsi sebagai pemecah kalimat berdasarkan tiap kata yang menyusunnya, yang disebut term atau token. Tokenizing dipecah berdasarkan spasi.

1. Filtering

Tahap ini memproses agar kata-kata yang tidak penting atau tidak bermakna dihapus untuk analisis sentimen. Contoh kata-kata tersebut adalah atau, yang, dengan, di, ke, dan tetapi.

1. Stemming

Tahap ini berfungsi mengubah kata berimbuhan pada tiap kata yang telah terseleksi menjadi kata dasar

* 1. **Term Frequency-Inverse Document Frequency (TF-IDF)**

Metode TF-IDF adalah perhitungan atau pembobotan kata melalui teknik tokenisasi, stopwords, dan stemming, dan frekuensi munculnya kata dalam dokumen yang diberikan menunjukkan pentingnya kata itu di dalam sebuah dokumen. Term Frequency (TF) adalah alat yang menghitung berapa sering sebuah kata muncul dalam dokumen. Inverse Document Frequency (IDF) adalah alat yang mengukur seberapa umum sebuah kata muncul di seluruh kumpulan dokumen. Semakin jarang kata muncul, semakin penting kata tersebut, Tokenisasi adalah proses memecah kalimat menjadi kata-kata, tanda baca, dan ekspresi bermakna lainnya sesuai dengan aturan Bahasa(Ramadhan & Ramadhan, 2022). Stopword adalah tahap mengambil kata penting dari hasil tokenizing dengan algoritma stopword removal (membuang kata yang kurang penting). Stopword merupakan kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan bag of words .Stemming adalah proses pengubahan sebuah kata ke bentuk dasarnya dengan menghilangkan imbuhan yang terdiri dari awalan, akhiran, awalan dan akhiran, dan sisipan .Dengan pembobotan ini, jika suatu term memiliki kemunculan yang tinggi di satu dokumen teks tapi memiliki kemunculan yang rendah di dokumen teks lain, maka bobotnya akan lebih tinggi. Namun, jika term tersebut sering muncul dalam dokumen teks lain, maka akan diberi bobot yang lebih rendah. Pada fase ini dilakukan eksekusi TF-IDF untuk memperoleh nilai TF berupa frekuensi kemunculan kata dari keseluruhan dokumen dan IDF berupa nilai setiap kata. Pemrosesan TF dan IDF dilakukan untuk setiap kelas, dan hasilnya dikalikan untuk mencari nilai bobot kata rumus TF-IDF yang digunakan adalah sebagai berikut(Gifari et al., 2022):

Dimana tf(i,j) adalah jumlah kemunculan term j dalam dokumen teks di ∈ D\*, dengan i = 1,2,3,....,N, df(j) adalah jumlah dokumen yang mengandung term j dari seluruh dokumen, dan N adalah jumlah seluruh dokumen yang ada di dalam dokumen. Berlandaskan rumus tersebut berapapun besarnya nilai tf(i,j), apabila N = df(j) akan didapatkan hasil 0 (nol) untuk perhitungan idf. Untuk itu dapat ditambahkan nilai 1 pada sisi idf, sehingga perhitungan untuk pembobotan dapat dilihat padarumus berikut :

Dimana :tf(i,j) = frekuensi kemunculan term j pada dokumen teksi = 1,2,3,....,NN = jumlah seluruh dokumen yang ada di koleksi dokumendf(j) = frekuensi dokumen yang mengandung term j dari semua koleksi dokumen(Keerthi Kumar et al., 2019)

* 1. **Support Vector Machine (SVM)**

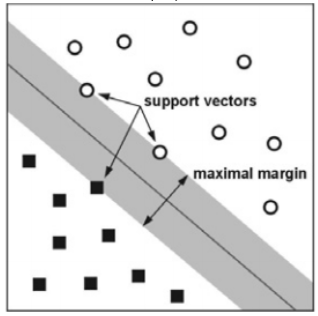
SVM adalah teknik untuk memprediksi klasifikasi dan regresi. SVM pertama kali diperkenalkan oleh Vapnik pada tahun 1992 sebagai seperangkat konsep unggul yang di bidang pattern recognition. SVM adalah algoritma pembelajaran mesin yang bekerja berdasarkan prinsip Structural Risk Minimization (SRM) dengan tujuan untuk menemukan hyperplane optimal yang memisahkan dua kelas dalam ruang input(Nuraliza et al., 2022). Berdasarkan Gambar 1, garis hyperplane berada pada tengah dari nilai maksimal margin dari data terdekat. Persamaan untuk mendapatkan hyperline:

A black and white image of a black and white object

Description automatically generated with medium confidence

Formula menghitung maksimal margin adalah:





Gambar 1 Support Vector Machine

1. **Hasil dan Pembahasan**

Hasil dan pembahasan akan dijabarkan dengan beberapa tahap yaitu:

* 1. **Persiapan**

Persiapan yang dilakukan adalah mempersiapkan alat dan dataset yang akan menjadi objek utama penelitian ini. Penelitian ini dikerjakan menggunakan processor 13-2310M 2.1Ghz dengan RAM 8GB dan menggunakan sistem operasi Windows 10 64-bit yang telah terinstall library python 3.11.5 Google Colab dan menggunakan library scikit-learn untuk membantu melakukan proses analisis ini. Dataset review IMDb yang memiliki 50.000 review.

|  |  |  |
| --- | --- | --- |
|  | Predicted | |
|  |  |
| Actual Negative | True Negative (TN) | False Positive (FP) |
| Actual Positive | False Negative(FN | True Positive (TP) |

Accuracy merupakan perbandingan antara semua

* 1. **Preprocessing Data**

Dalam tahap ini data tersebut akan diproses agar dimensi text menjadi lebih sederhana. Contoh text yang akan memasuki preprocessing: “\*\*\*When I first see it, I am too young around 15, to understand all the things the movie what to show us. The film gave me a very different feeling. Hope is a dangerous thing, is the root of mental distress.”. Hasil dapat dilihat pada tabel 3 berikut.

|  |  |
| --- | --- |
| Tahap | Output |
| Cleaning dan Case- Folding | when i first see it i am too young around to understand all the things the movie what to show us the film gave me a very different feeling hope is a dangerous thing is the root of mental distress |
| Stopword removal | first see young around understand things movie show us film gave different feeling hope dangerous thing root mental distress |
| Stemming | first see young around understand thing movi show us film gave differ feel hope danger thing root mental distress |

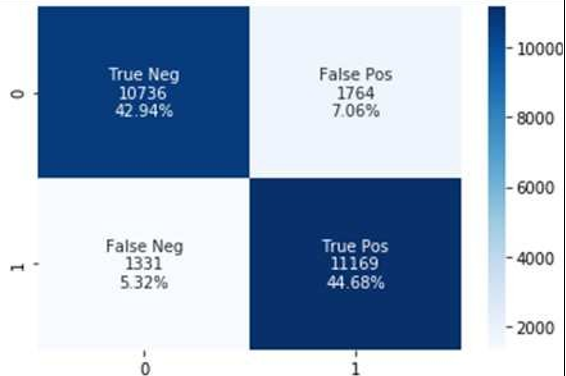
Dapat dilihat bahwa melakukan preprocessing data dapat memangkas data yang berawal besar menjadi lebih kecil dan lebih bermakna. Karakter pada data yang belum diprocessing terhitung sebanyak 207 karakter kamudian setelah melewati proses Cleaning dan Case-Folding karakter data menjadi 196, data tersebut memasuki tahap stopwords removal dan berhasil memangkas karakter menjadi 124 karakter, tahap terakhir adalah stemming data tersebut dan berhasil memangkas karakter menjadi 113. Dalam hal ini preprocessing data lebih tepatnya normalisasi data yang berawal 207 karakter menjadi hanya 113 karakter yang dapat diproses untuk klasifikasi.

* 1. **Klasifikasi**

Tahap ini melakukan eksperimen yang dengan melakukan klasifikasi text menggunakan dalam dataset menggunakan algoritma SVM adalah Linier SVM. Untuk SVM linier menggunakan parameter C=0.01 dan tolerasi 0.001 selain itu menggunakan parameter default dari scikit-learn library.

|  |  |
| --- | --- |
| 25000 data | SVM Linier |
| Accuracy | 0.8762000000000000 |
| Precision | 0.8636047320807237 |
| Recall | 0.8935200000000000 |
| F1-Score | 0.8783077104549208 |

Hasil dari Accuracy, Precision, Recall dan F1-Score didapat menggunakan Confusion Matrix. Setiap metode yang digunakan memiliki Confusion matrix sendiri dengan value yang berbeda-beda sesuai dengan kinerja metode yang digunakan. Eksperimen ini menggunakan library Seaborn untuk membuat diagram Confusion Matrix menjadi lebih mudah. Berikut tabel Confusion Matrix.



Dari hasil diatas dapat dilihat bahwa nilai dari *Confusion matrix,* model ini menghasilkan 10736 prediksi negatif yang benar (True Negatives) yang merupakan 42.94% dari total prediksi. Sebaliknya, model juga menghasilkan 1764 prediksi positif yang salah (False Positives), yang merupakan 7.06% dari total prediksi, menunjukkan kesalahan tipe I. Selain itu, terdapat 1331 prediksi negatif yang salah (False Negatives), yang merupakan 5.32% dari total prediksi, menunjukkan kesalahan tipe II. Terakhir, model menghasilkan 11169 prediksi positif yang benar (True Positives), yang merupakan 44.68% dari total prediksi. Kesimpulannya, model ini lebih sering membuat prediksi yang benar daripada yang salah, namun ada ruang untuk peningkatan dalam mengurangi kesalahan tipe I dan tipe II.

1. **Kesimpulan dan Saran**

Berdasarkan hasil yang diperoleh pada penelitian ini, terbukti bahwa tujuan untuk menganalisis sentimen pada komentar pengguna situs web IMDB dengan aspek rating dapat dilakukan. Model klasifikasi SVM yang digunakan pada penelitian ini juga mampu menghasilkan akurasi sebesar 87%, presisi sebesar 86%, dan recall sebesar 89%. Hal ini sudah bisa dikatakan baik.

**Daftar Pustaka:**

Gifari, O. I., Adha, M., Rifky Hendrawan, I., Freddy, F., & Durrand, S. (2022). Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine. *JIFOTECH (JOURNAL OF INFORMATION TECHNOLOGY*, *2*(1).

Kardila, Y. (2023). Analisis Sentimen Review Pengguna Website IMDB Menggunakan Klasifikasi Naïve Bayes. *E-Proceeding of Engineering*, *10*, 3156.

Keerthi Kumar, H. M., Harish, B. S., & Darshan, H. K. (2019). Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method. *International Journal of Interactive Multimedia and Artificial Intelligence*, *5*(5), 109–114. https://doi.org/10.9781/ijimai.2018.12.005

Nur Cahyo, D. D., Farasalsabila, F., Lestari, V. B., Hanafi, Lestari, T., Al Islami, F. R., & Maulana, M. A. (2023). Sentiment Analysis for IMDb Movie Review Using Support Vector Machine (SVM) Method. *Inform : Jurnal Ilmiah Bidang Teknologi Informasi Dan Komunikasi*, *8*(2), 90–95. https://doi.org/10.25139/inform.v8i2.5700

Nuraliza, H., Pratiwi, O. N., & Hamami, F. (2022). Analisis Sentimen IMBd Film Review Dataset Menggunakan Support Vector Machine (SVM) dan Seleksi Feature Importance. In *Jurnal Mirai Manajemen* (Vol. 7, Issue 1).

Ramadhan, N. G., & Ramadhan, T. I. (2022). Analysis Sentiment Based on IMDB Aspects from Movie Reviews using SVM. *Sinkron*, *7*(1), 39–45. https://doi.org/10.33395/sinkron.v7i1.11204

##### Lampiran

Makalah dapat dilengkapi dengan lampiran dengan tidak melebihi batas maksimal jumlah halaman