p–ISSN: 2723 - 6609 e-ISSN : 2745-5254
Vol. 4, No. 4, April 2023 http://jist.publikasiindonesia.id/

SENTIMENT ANALYSIS TASKS DENGAN METODE KLASIFIKASI MACHINE LEARNING (STUDY KASUS REPOSITORI REVIEW MOVIE) DENGAN WEKA

Damar Wicaksono[1*], Imam Adi Nata[2]
Universitas Tidar Magelang, Indonesia
Email: damar@untidar.ac.id[1*], imamadinata@untidar.ac.id[2]

*Correspondence

INFO ARTIKEL

Diterima : 07-04-2023
Direvisi : 15-04-2023
Disetujui : 17-04-2023

Kata kunci: review film; text mining; klasifikasi; peminatan; opini publik.

ABSTRAK

Sebuah proses pengambilan informasi berkualitas tinggi dari sebuah teks yang akan diproses disebut sebagai text mining. Informasi tersebut seringkali didapatkan melalui proses pengenalan pola dan berbagai kemungkinan melalui cara seperti pembelajaran berpola statistik. Adapun pada saat ini, jumlah arus informasi yang sangat besar telah tersedia dalam dokumen online. Penambangan berupa data teks dimulai dari kebutuhan memproses dalam bentuk teks data tidak terstruktur. Penambangan data teks merupakan turunan penambangan data dan oleh karenanya memiliki banyak similaritas metode implementasinya. Sebagai bagian dari upaya untuk lebih mengatur informasi tersebut bagi pengguna, maka dilakukan penyelidikan masalah agar teks dapat terkategorisasi secara otomatis. Makalah ini membahas tentang pemrosesan text mining menggunakan berbagai macam metode klasifikasi yang digunakan untuk dapat mengetahui seberapa besar tingkat peminatan pemirsa terhadap sebuah acara review film. Dataset digunakan pada penelitian tersebut diambil dari sebuah data berita. Dataset yang disediakan terdiri dua direktori yakni direktori pos dan neg pada direktori data_review_film dalam penelitian ini. Data ini merupakan data opini publik tentang sebuah proses review film yang kemudian akan dilakukan prediksi berapakah jumlah opini tersebut yang merupakan kombinasi kemungkinan dari review positif dan negatif. Pada bagian pembahasan dilakukan perbandingan untuk masing-masing proses klasifikasi yang dilakukan. Sehingga dalam penelitian ini, klasifikasi Naïve Bayes merupakan algoritma paling efektif dibandingkan dengan algoritma lainnya.

Keywords: *movie review; text mining; classification; interest; public opinion.*

ABSTRACT

A process of extracting high quality information from a text to be processed is known as text mining. Such information is often obtained through forecasting patterns and trends through tools such as statistical pattern learning. As for now, a very large amount of information flow has been available in online documents. Mining in the form of text data starts from the need to process text in the form of unstructured data. Text data mining is a derivative of data mining and therefore has many similarities in its implementation methods. As part of an effort to better organize this information for users, a problem investigation was carried out so that the text could be automatically categorized. This paper discusses the processing of text mining using various classification methods that are used to find out how much the audience's level of interest in a film review program is. The dataset in this study was taken from a news data. Datasets provided consists of two directories, namely the pos and neg directories in the data_review_film directory in this study. This data is public opinion data about a film review process which will then be predicted how many opinions will be which is a possible combination of positive and

negative reviews. In the discussion section, a comparison is made for each of the classification methods used. So in this study, the Naïve Bayes classification is the most effective algorithm compared to other algorithms.

Pendahuluan

Bahasa adalah sebuah alat yang digunakan sebagai media komunikasi dan untuk memindahkan informasi (Mailani, Nuraeni, Syakila, & Lazuardi, 2022). Bahasa juga juga dilihat sebagai sarana mengekspresikan sentimen dan emosi. Proses penentuan apakah positif, negatif atau netral pada isi suatu dataset berupa teks berupa dokumen, kalimat dan paragraf disebut sebagai *sentiment analysis* (Darwis, Siskawati, & Abidin, 2021). Analisis sentimen di bidang penambangan pendapat pengguna pada produk, ulasan politik, ulasan film kini menjadi lebih populer. Produser dan pembuat film melalui media sosial dan IMD dapat mengetahui review, pandangan dan pemikiran dari para pemirsanya (Buniar, Utsalinah, & Wahyuningsih, 2022).

Sedangkan *sentiment analysis* untuk film merupakan penilaian isi dari *review* film yang diproses dengan penerapan *natural language processing* dan analisis teks dengan tujuan untuk proses identifikasi dan ektraksi dari dari informasi yang bersifat subjektif dari sebuah teks (Hussein, 2018). Proses ini meliputi memahami dan mengolah data tekstual secara otomatis guna mendapat informasi *sentiment* terkandung dalam suatu kalimat pendapat/opini. Proses tersebut dilakukan agar dapat memahani kecenderungan terhadap sebuah masalah apakah cenderung berpandangan atau beropini positif atau negatif (Nurzahputra & Muslim, 2016). Penerapan klasifikasi ini menjadi kalimat positif maupun negatif dapat dilakukan setelah simplifikasi data subjek yang digunakan untuk mereduksi fitur agar menghindari banyaknya dimensi dipakai pada saat proses klasifikasi ini berlangsung (Savitri, Rahman, Venyutzky & Rakhmawati, 2021).

Proses review film memiliki ukuran dataset besar pada level data latih maupun pada data uji. Ukuran dimensi dan berbagai fitur berlebih dapat meningkatkan ruang pencarian semakin tinggi, yang menyebabkan sulitnya dalam 1) kompleksitas memproses data dan 2) akan menurunkan nilai kinerja serta 3) membuat data yang diproses menjadi tidak konsisten. Analisis dan penambangan juga membutuhkan waktu yang relatif lebih lama. Selanjutnya pengurangan dimensi ini bertujuan untuk mengurangi level kompleksitas dimensi data, dan nantinya akan meningkatkan kinerja *machine learning* serta mengeliminasi dari ekstraksi fitur yang tidak dibutuhkan (Pristiyanti, Fauzi, & Muflikhah, 2018).

Pada kasus tertentu, dikemukakan sebuah review produk apakah bernilai positif atau negatif. Label artikel ini dengan sentimen mereka akan memberikan ringkasan singkat bagi pembaca. Tidak dapat dipungkiri bahwa label ini merupakan bagian dari daya tarik dan menjadi nilai tambah bagi website seperti www.rottentomatoes.com, yang keduanya label ulasan film yang tidak mengandung indikator peringkat yang dinyatakan secara eksplisit dan melakukan proses normalisasi untuk skema

pemeringakatan yang berbeda yang biasa digunakan bagi para reviewernya. Klasifikasi Sentimen ini juga akan membantu dalam aplikasi kecerdasan bisnis (misalnya sistem pada perangkat lunak MindfulEye oleh Lexant system1) dan sistem untuk rekomendasi (Rintyarna, 2022).

Secara umum, apapun bentuk respon survei yang diberikan dalam format bahasa alami dapat diproses dengan menggunakan kategorisasi sentimen. Selain itu, terdapat juga aplikasi yang potensial untuk penyaringan pesan, misalnya satu mungkin dapat menggunakan informasi sentimen untuk mengenali dan membuang konten tertentu yang dikehendaki (Balan & Mathew, 2015).

Dalam makalah ini menjelaskan bagaimana sebuah efektivitas dari penerapan teknik machine learning untuk masalah klasifikasi sentimen. Sebuah tantangan pada masalah ini yang tampak membedakan dari klasifikasi berdasarkan topik tradisional bahwa sementara topik yang didapatkan sering diidentifikasi oleh kata kunci saja, sentimen dapat dinyatakan dengan cara yang berasumsi bahwa sentimen memerlukan pemahaman yang lebih dari klasifikasi biasa berdasarkan topik yang telah didapatkan. Jadi, selain menyajikan hasil juga dapat dilakukan analisis masalah untuk mendapatkan pemahaman yang lebih baik (Rintyarna, 2016).

Sedangkan tajuk masalah dibahas dalam paper ini yakni bagaimana membuat proses klasifikasi dari masalah dokumen dianggap tidak hanya berdasarkan pada topik saja, akan tetapi oleh parameter sentimen secara keseluruhan dan menentukan apakah tinjauan tersebut positif atau negatif dengan menggunakan review film sebagai data. Percobaan ini melakukan klasifikasi raw data review film menggunakan perbandingan antara metode dari 4 algoritma tersebut diantaranya : Naïve Bayes, k-Nearest Neighbors algorithm, TreeJ48 dan ADTree dengan perangkat bantu WEKA. Dataset berupa berita merupakan dataset yang diambil melalui website yakni “Sentiment Polarity Dataset version 2.0” (http://www.cs.cornell.edu/People/pabo/movie-review-data). Didalamnya terdapat dua kelas diantaranya *review* positif dan *review* negatif. Data tersebut berupa variasi opini/pendapat dari analisis sebuah task yang dibagi menjadi dua oleh reviewer berupa kalimat yang bersifat subjektif dan objektif.

Metode Penelitian

Metode digunakan dalam penelitian ini merupakan metode klasifikasi. Dimana klasifikasi merupakan tindakan untuk memberi kelompok pada setiap kondisi salah satunya *class attribute*. Metode tersebut dapat ditemukan pada sebuah model yang dapat menerjemahkan atribut kelas tersebut sebagai fungsi dari masukan. Sedangkan dataset digunakan pada penelitian ini berasal dari sebuah data berita. Dataset yang disediakan terdiri dua direktori yakni pos dan neg pada direktori data_review_film dalam penelitian ini. Data ini merupakan data opini publik tentang sebuah proses review film (movie) yang kemudian akan dilakukan prediksi berapakah jumlah opini tersebut yang merupakan kombinasi kemungkinan dari review positif dan negatif. Dataset berita berasal dari dataset diambil dari website dan telah ditentukan disebut sebagai “Sentiment Polarity Dataset version 2.0” dimana dataset tersebut terdiri dari dua jenis

kelas: review positif dan review negatif. Data tersebut berupa variasi opini/pendapat dari analisis sebuah task yang dibagi menjadi dua oleh reviewer berupa kalimat subjektif dan objektif. Dataset yaitu data_review_film terdiri atas 2000 file teks dibagi menjadi dua sub-kategori direktori (pos dan neg dari nilai kelas).

1. Pra pengolahan

Tahapan ini digunakan sebagai cara untuk memperbaiki data sebelum dilakukan proses input *machine learning*. Sedangkan capaiannya untuk menghasilkan nilai performansi yang lebih baik (Alita, Fernando, & Sulistiani, 2020). Penyebab data yang kurang baik adalah Noisy yang berisi kesalahan atau nilai *outlier* menyimpang dan tidak konsisten berupa ketidakcocokan dalam penggunaan kode atau nama. Sehingga perlu dilakukan tahapan untuk dataset pada penelitian ini diantaranya:

1. Proses analisis untuk mencari apakah ada atribut *outlier* dan memperbaikinya.
2. Filtering terhadap satu atau lebih atribut untuk mendapatkan performansi klasifikasi yang lebih baik.

Tahap preprocessing dalam penelitian ini menggunakan teknik *data-cleaning*. Menggunakan teknik tersebut, maka dalam tahap ini dilakukan proses untuk menghilangkan nilai-nilai data tertentu yang dianggap salah, memperbaiki volatilitas data dan melakukan verfikasi data yang tidak konsisten. Sedangkan identifikasi dan pemilihan atribut dan proses diskritisasi nilai (attribute identification and selection) digunakan untuk membatasi agar komputasi lebih cepat dan meningkatkan performansi pada pemrosesan dataset tersebut.

2. Klasifikasi

a. Tahap awal dalam proses pelatihan algoritma klasifikasi pada dataset data_review_film adalah menyiapkan data dengan teliti dan hati-hati.
b. Sebelum melakukan pelatihan dengan algoritma pada tools Weka, data_review_film harus dipersiapkan terlebih dahulu agar dapat dimengerti oleh tools tersebut.
c. Persiapan data yang cermat dan tepat merupakan tahap pertama dalam pelatihan algoritma klasifikasi pada dataset data_review_film dengan menggunakan tools Weka.

Metode Naïve Bayes

Naïve Bayes merupakan metode yang erat hubungannya dengan klasifikasi, korelasi hipotesis, dan bukti klasifikasi. Jika kita memiliki vektor masukkan yang berisi fitur A dan label kelas B, maka notasi P(A|B) digunakan untuk menunjukkan probabilitas label kelas B yang didapatkan setelah fitur-fitur A diamati. Notasi ini juga dikenal dengan sebutan probabilitas posterior untuk B, sedangkan P(B) disebut probabilitas prior untuk B. Dalam banyak kasus, metode Bayes mudah dihitung untuk fitur kategorikal seperti dalam klasifikasi tanaman dengan fitur "penutup lapisan kulit" yang memiliki nilai {epidermis, meristematik, kolenkim, dan sklerenkim} atau kasus fitur "jenis bagian badan" dengan nilai {daun, ranting, batang, dan akar}.

Notasi P(A|B) dalam metode Naïve Bayes digunakan untuk menghitung probabilitas label kelas B setelah fitur-fitur A diamati. Probabilitas ini dikenal sebagai

probabilitas posterior untuk B, sedangkan P(B) merupakan probabilitas prior untuk B. Metode Naïve Bayes banyak digunakan dalam klasifikasi, korelasi hipotesis, dan bukti klasifikasi, terutama untuk fitur kategorikal seperti dalam klasifikasi tanaman dengan fitur "penutup lapisan kulit" yang memiliki nilai {epidermis, meristematik, kolenkim, dan sklerenkim} atau fitur "jenis bagian badan" dengan nilai {daun, ranting, batang, dan akar}.

Metode Naïve Bayes sering digunakan dalam klasifikasi, korelasi hipotesis, dan bukti klasifikasi karena sangat terkait dengan probabilitas posterior dan prior. Notasi P(A|B) digunakan untuk menghitung probabilitas label kelas B setelah fitur-fitur A diamati, sementara P(B) merupakan probabilitas prior untuk B. Dalam kasus fitur kategorikal seperti dalam klasifikasi tanaman dengan fitur "penutup lapisan kulit" yang memiliki nilai {epidermis, meristematik, kolenkim, dan sklerenkim} atau fitur "jenis bagian badan" dengan nilai {daun, ranting, batang, dan akar}, metode Naïve Bayes sangat mudah dihitung.
Formulasi Naïve Bayes sebagai fungsi klasifikasi dapat didefinisikan sebagai berikut:

$$\mathrm{P(B|A)} = \frac{P(B)\Pi_{\mathrm{i=1}}^{\mathrm{q}}\mathrm{P(A_i|B)}}{\mathrm{P(x)}}$$

P(B|A) merupakan probabilitas data dengan vektor A untuk kelas B. P(B) adalah probabilitas awal kelas dengan asumsi B. Kemudian P(B) $\Pi_{\mathrm{i=1}}^{\mathrm{q}}$ $^{\mathrm{q}}$ $\mathrm{P(A_i \mid B)}$ adalah probabilitas independen kelas B dari semua fitur dalam vektor A. Nilai P(A) selalu konstan sehingga dalam perhitungan prediksi tinggal dihitung bagian P(B) $\Pi_{\mathrm{i=1}}^{\mathrm{q}}\mathrm{P(A_i \mid B)}$ dengan mengambil nilai terbesar sebagai kelas yang terpilih sebagai hasil klasifikasi. Sedangkan probabilitas mandiri yang digunakan pada persamaan notasi
$\Pi_{\mathrm{i=1}}^{\mathrm{q}}$ $\mathrm{P(A_i \mid B)}$ merupakan pengaruh semua bagian fitur dari data terhadap setiap kelas B, yang dapat dinotasikan sebagai:

$$\mathrm{P(A|B = y)} = \Pi_{\mathrm{i=1}}^{\mathrm{q}}\mathrm{P(A_i|B = y)}$$

Untuk tiap set fitur A = {A1, A2, A3…, Aq} terdiri atas q atribut (q buah dimensi).

Klasifikasi dengan Naïve Bayes dapat diimplementasikan didasarkan oleh teori probabilitas yang memandang semua fitur dari data sebagai bukti dalam probabilitas dan memiliki karakteristik sebagai berikut:

1. Metode Naïve Bayes bersifat *robust* terhadap data-data yang terpisah dan biasanya adalah data dengan karakteristik berbeda (*outliner*).
2. Metode ini mampu menangani nilai atribut yang tidak sesuai dengan menghiraukan data latih selama proses pembangunan model dan prediksi.
3. Cukup handal dalam menghadapi data dengan atribut yang tidak relevan.
4. Bagiam Atribut mempunyai korelasi bisa menurunkan kinerja metode klasifikasi ini dikarenakan asumsi independensi bagi atribut yang sudah tidak ada.

Metode KNN (IBK)

KNN atau IBK adalah sebuah metode klasifikasi dalam Weka yang memanfaatkan data dari proses pembelajaran dengan mencari jarak terdekat antara objek yang akan diklasifikasikan dengan objek yang sudah ada pada memori. Metode supervised learning mengasumsikan bahwa hasil klasifikasi instance baru adalah mayoritas kategori pada data training. Model klasifikasi ini tidak memerlukan kategori model tertentu untuk dipasangkan, hanya menggunakan memori yang sudah tersedia. Pada titik query, model akan mencari k-objek terdekat dengan instance yang baru dan memilih kategori mayoritas sebagai prediksi untuk instance baru tersebut.

Metode K-Nearest Neighbor (KNN) merupakan algoritma yang simpel, yang bekerja dengan cara mencari jarak terdekat antara query instance dan training sample untuk menentukan tetangga terdekatnya. Proses pelatihan diilustrasikan dalam ruang dengan banyak dimensi, di mana setiap dimensi merepresentasikan fitur dari data yang diamati. Ini menghasilkan ruang yang terbagi menjadi beberapa bagian berdasarkan klasifikasi contoh pelatihan. Titik dalam ruang ini akan diklasifikasikan ke dalam kelas tertentu jika kelas tersebut adalah kelas yang paling sering muncul pada k tetangga terdekat dari titik tersebut. Biasanya, jarak antara tetangga dihitung menggunakan Euclidean Distance. Jarak Euclidean paling sering digunakan menghitung jarak. Jarak euclidean berfungsi menguji ukuran yang bisa digunakan sebagai interpretasi kedekatan jarak antara dua obyek. yang direpresentasikan sebagai berikut: (Pristiyanti, Fauzi, & Muflikhah, 2018)

$$D(a,b) = \sqrt{\sum_{i=1}^{n} (b_i - a_i)^2}$$

Ket :
matriks D (a, b) merupakan skalar jarak diantara kedua vektor a dan vektor b dari matriks dengan ukuran dimensi tertentu.

Nilai D yang semakin besar ini akan menjadikan semakin jauhnya kemiripan antara kedua *instance* dan demikian berlaku juga sebaliknya. Nilai k terbaik untuk algoritma ini tergantung pada data yang telah disaring.

Sehingga secara umum, nilai k yang tinggi akan mengurangi derau pada proses klasifikasi, dan membuat batasan diantara setiap proses klasifikasi menjadi semakin kabur. Nilai k yang terpilihh didapatkan dengan optimasi parameter yang ada, misalnya dengan menggunakan proses validasi silang. Diaman pada kasus tertentu, klasifikasi ditentukan berdasarkan data latih yang paling dekat (dengan kata lain, nilai k = 1) disebut sebagai algoritma tetangga terdekat. Ketepatan algoritma ini sangat dipengaruhi oleh ada ataupun tidaknya fitur yang tidak relevan. Riset terhadap algoritma inipun secara mayoritas telah dijelaskan bagaimana cara untuk menyeleksi dan memberikan

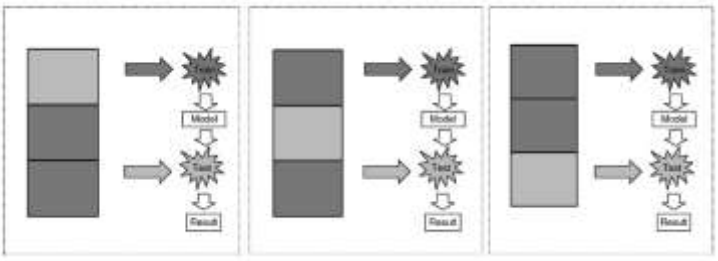
bobot tertentu dari fitur agar performansi klasifikasi agar menjadi lebih baik. Berikut ini merupakan langkah untuk menghitung metode K-Nearest Neighbor:

1. Memberi nilai dari parameter K (dengan jumlah tetangga paling dekat).
2. Melakukan kalkulasi dari kuadrat jarak Euclidian (*query instance*) dari masing-masing objek dari contoh data yang diproses.
3. Mengurutkan objek-objek kedalam kategori label dan memiliki jarak Euclidian terpendek.
4. Mengumpulkan nilai kategori dari B (klasifikasi tetangga terdekat)
5. Dengan menggunakan kategori tetangga terdekat yang paling dominan sehingga dapat diramalkan nilai *query instance* terhitung.

Decision Tree (TreeJ48 dan ADTree)

Algoritma ADTree dan TreeJ48 memiliki pendekatan yang sama dalam membentuk decision tree dari atas ke bawah (top-down). Decision tree digunakan untuk menyajikan algoritma dengan pernyataan kontrol bersyarat yang menghasilkan keputusan yang menguntungkan. Setiap cabang pada pohon keputusan mewakili hasil untuk atribut dan jalur dari daun ke akar mewakili aturan klasifikasi. Metode ini digunakan untuk mengklasifikasikan data dengan mempertimbangkan nilai atribut. Dalam validasi silang, data pengujian dibagi secara acak menjadi k himpunan dengan ukuran yang sama.

Dalam pengujian dan pelatihan model, dilakukan sebanyak k kali iterasi. Pada setiap iterasi ke-i, bagian data Di digunakan sebagai data uji, sedangkan bagian data yang tersisa digunakan sebagai data latih untuk membangun model pada iterasi tersebut. Sebagai contoh, pada iterasi pertama, bagian data kedua hingga ke-k digunakan sebagai data latih untuk membangun model pertama, dan model tersebut diuji pada bagian data pertama. Iterasi kedua dilakukan dengan menggunakan bagian data pertama dan data bagian ketiga hingga ke-k sebagai data latih dan diuji pada bagian data kedua, dan seterusnya hingga selesai seperti yang dijelaskan pada Gambar 1. Dalam penelitian ini, digunakan validasi silang 10 kali lipatan untuk menguji dan melatih model.



Gambar 1 Ilustrasi dari *3-fold cross validation*

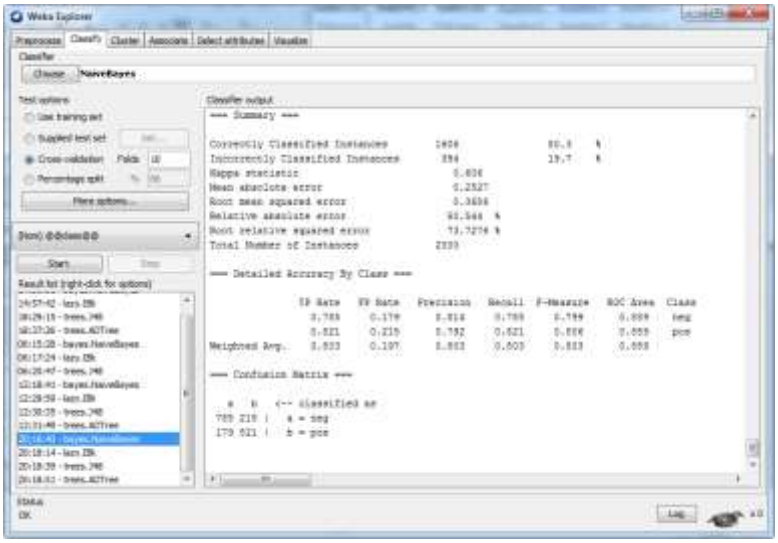
Evaluasi

Pada tahapan akhir dalam penelitian ini, dilakukan perlunya mengevaluasi performansi dari metode atau algoritma klasifikasi yang sudah dipakai dan menunjukkan beberapa indikator performansi yang akan dijelaskan pada bagian pembahasan masing-masing metode yang dipakai.

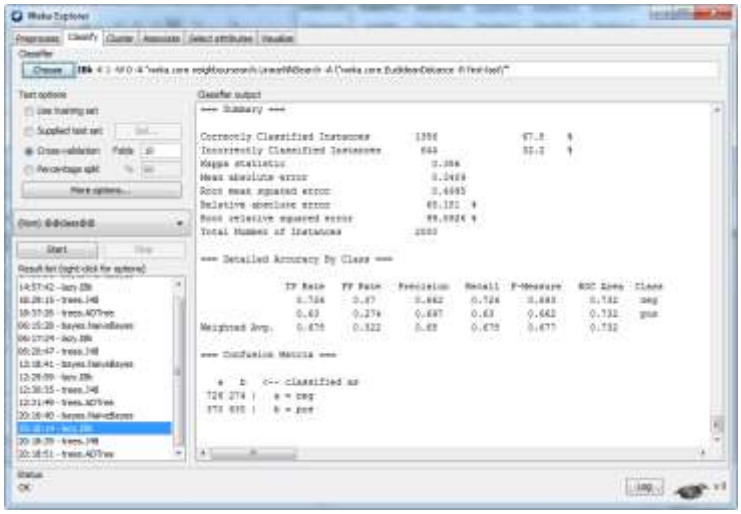
Hasil dan Pembahasan

Proses klasifikasi adalah proses untuk klasifikasi atribut pada teks yang memliki 1000 atribut untuk review opini positif dan negative dan direduksi hingga 53 buah. Proses klasifikasi ini dilakukan dengan menggunakan 4 algoritma text mining untuk mengetahui algoritma mana yang memiliki tingkat akurasi performansi yang paling tinggi. 4 algoritma tersebut adalah Naïve Bayes, k-Nearest Neighbors algorithm (KNN), TreeJ48 dan ADTree. Proses klasifikasi ini dilakukan dengan menggunakan Weka dengan menggunakan percentage split sebesar 66% bisa dilihat pada Gambar 2 hingga 13 berikut ini.

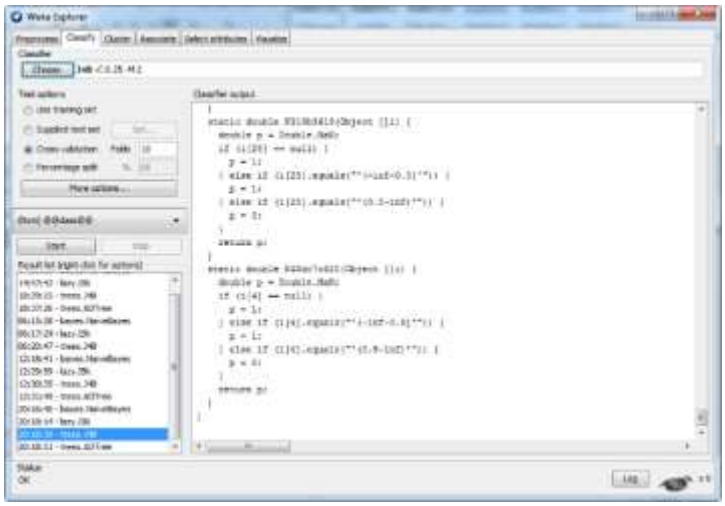
Proses klasifikasi ini dilakukan dengan menggunakan Weka dengan menggunakan *percentage split* sebesar 66% bisa dilihat pada Gambar 2 hingga Gambar 5 berikut ini.



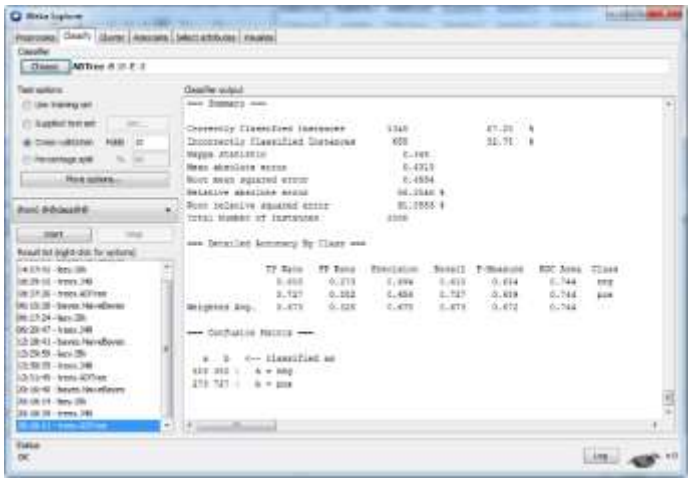
Gambar 2 Klasifikasi data dengan metode Naïve Bayes



Gambar 3 Klasifikasi data dengan metode k-Nearest Neighbors algorithm (KNN)



Gambar 4 Klasifikasi data dengan metode Tree J48



Gambar 5 Klasifikasi data dengan ADTree

Berdasarkan dari percobaan yang sudah dilakukan diatas, dapat digali informasi seperti yang ditampilkan dalam Tabel seperti dibawah ini.
Tabel Confusion Matrix dengan menggunakan Model Test : 10-fold cross-validation

Tabel 1
Hasil Pengujian menggunakan metode Naïve Bayes

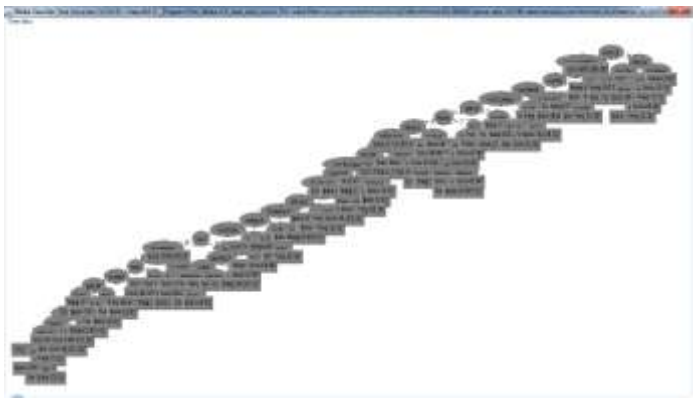
Hasil Pengujian Naïve Bayes	Review untuk film		
	Positif	Negatif	
Positif	785	215	1000
Negatif	179	821	1000
	964	1036	2000

Tabel 2
Hasil Pengujian menggunakan metode k-Nearest Neighbors algorithm

Hasil Pengujian Ibk/KNN	Review untuk film		
	Positif	Negatif	
Positif	726	274	1000
Negatif	370	630	1000
	1096	904	2000

Tabel 3
Hasil Pengujian menggunakan metode TreeJ48

Hasil Pengujian trees. J48	Review untuk film		
	Positif	Negatif	
Positif	750	250	1000
Negatif	262	738	1000
	1012	988	2000



Sedangkan ditemukan jumlah cabang dalam *tree decision*

Jumlah daun terbentuk:	100
Ukuran pohon:	199

Tabel 4
Hasil Pengujian menggunakan metode *ADTree*

Hasil Pengujian ADTree	Review untuk film		
	Positif	Negatif	
Positif	618	382	1000
Negatif	273	727	1000
	891	1109	2000

Tabel 5
Hasil Summary secara keseluruhan untuk masing-masing metode

Metode	**Naïve Bayes**	**Ibk/KNN**	**Trees. J48**	**ADTree**
Correctly Classified Instances	1606 (80.3 %)	1356 (67.8 %)	1488 (74.4 %)	1345 (67.25 %)
Incorrectly Classified Instances	394 (19.7 %)	644 (32.2 %)	512 (25.6 %)	655 (32.75 %)
Kappa statistic	0.606	0.356	0.488	0.345
Mean absolute error	0.2527	0.3409	0.319	0.4313
Root mean squared error	0.3686	0.4995	0.4469	0.4554
Metode	**Naïve Bayes**	**Ibk/KNN**	**Trees. J48**	**ADTree**
Relative absolute error	50.54%	68.18%	63.80%	86.25%
Root relative squared error	73.73%	99.89%	89.38%	91.09%
Total Number of Instances	2000	2000	2000	2000

Melalui Tabel 1 tersebut diatas dapat diterjemahkan bahwa jumlah data uji untuk reviewer film yang yang diduga memiliki jumlah opini positif 1000 dimana 785 dari jumlah pasien yang ada (true-positive/TP) didapatkan dengan benar memberikan opini positif, sedangkan 215 review film yang yang diduga salah dalam identifikasi (false-positive/FP) oleh pengklasifikasi Naïve Bayes dimana kondisi sebenarnya review film tersebut tidak memiliki opini positif. Sedangkan pada pengujian pada review film yang diduga memiliki opini negatif menunjukkan 179 reviewer (true-negative/TN) dikenali

dengan benar tidak memiliki opini positif, sebaliknya terdapat 821 reviewer (false-negative/FN) salah dikenali sebagai reviewer film yang yang tidak memiliki opini positif.

Kemudian dari Tabel 2 diatas dapat dapat dijelaskan bahwa jumlah data pengujian untuk reviewer film yang yang diduga memiliki jumlah opini positif 1000 dimana 726 pasien (true-positive/TP) terdeteksi dengan benar memberikan opini positif, sedangkan 274 review film yang yang diduga salah diidentifikasi (false-positive/FP) oleh pengklasifikasi Ibk/KNN dimana kondisi sebenarnya review film tersebut tidak memiliki opini positif. Sedangkan pada pengujian pada review film yang diduga memiliki opini negatif menunjukkan 370 reviewer (true-negative/TN) dikenali dengan benar tidak memiliki opini positif, sebaliknya terdapat 630 reviewer (false-negative/FN) salah dikenali sebagai reviewer film yang yang tidak memiliki opini positif.

Dari Tabel 3 diatas dapat dijelaskan bahwa jumlah data pengujian untuk reviewer film yang yang diduga memiliki jumlah opini positif 1000 dimana 750 pasien (true-positive/TP) terdeteksi dengan benar memberikan opini positif, sedangkan 250 review film yang yang diduga salah diidentifikasi (false-positive/FP) oleh pengklasifikasi Naïve Bayes dimana kondisi sebenarnya review film tersebut tidak memiliki opini positif. Sedangkan pada pengujian pada review film yang diduga memiliki opini negatif menunjukkan 262 reviewer (true-negative/TN) dikenali dengan benar tidak memiliki opini positif, sebaliknya terdapat 738 reviewer (false-negative/FN) salah dikenali sebagai reviewer film yang yang tidak memiliki opini positif.

Dari Tabel 4 diatas dapat dijelaskan bahwa jumlah data pengujian untuk reviewer film yang yang diduga memiliki jumlah opini positif 1000 dimana 618 pasien (true-positive/TP) terdeteksi dengan benar memberikan opini positif, sedangkan 382 review film yang yang diduga salah diidentifikasi (false-positive/FP) oleh pengklasifikasi Naïve Bayes dimana kondisi sebenarnya review film tersebut tidak memiliki opini positif. Sedangkan pada pengujian pada review film yang diduga memiliki opini negatif menunjukkan 273 reviewer (true-negative/TN) dikenali dengan benar tidak memiliki opini positif, sebaliknya terdapat 727 reviewer (false-negative/FN) salah dikenali sebagai reviewer film yang yang tidak memiliki opini positif.

Dari matrik confusion pada Tabel sebelumnya dapat dilihat pada Tabel 5 dapat dihitung akurasi dari pengklasifikasi Naïve Bayes mencapai 80.3 %, Tabel 2 dari pengklasifikasi KNN mencapai 67.8 %, Tabel 3 dari pengklasifikasi decision tree J48 mencapai 74.4 % dan Tabel 4 menggunakan ADTree sebesar 67.25 %. Besarnya kesalahan yang menyebabkan penurunan akurasi terjadi pada kondisi false-positive.

Kesimpulan

Ekstraksi informasi menggunakan text mining dari dataset review movie apakah bernilai positif atau negative sangat efektif sebagai sistem analisis sentimen bagi praktisi bisni dari data mining adalah untuk mendapatkan pola informasi yang tersimpan dalam suatu teks yang dapat digunakan untuk pengolahan selanjutnya dan sebagai bahan pendukung keputusan dalam sistem penilaian sebuah produk,

Melalui penelitian dan percobaan yang sudah dilakukan, didapatkan beberapa analisis sebagai berikut:

a. Pra pengolahan pada dataset di awal sebelum melakukan proses klasifikasi dengan algoritma penambangan teks mempengaruhi performansi dari algoritma tersebut. Pada penelitian ini, teknik *data-cleaning oulier*, *stop-words removal*, *detection attribute selection* dan *discretize* cukup efektif dalam proses menghilangkan data redundant yang tidak diperlukan.
b. Dari keempat algoritma yang dilakukan percobaan didapatkan bahwa metode klasifikasi menggunakan Naïve Bayes menghasilkan tingkat akurasi paling tinggi diantara algoritma lainnya. Sedangkan ADTree memliki tingkat akurasi yang paling rendah sebesar 67,25%.
c. Kualitas data yang tinggi juga dapat mempengaruhi performansi penambangan data dalam menentukan atribut dari masing-masing kelas.

Bibliografi

Alita, Debby, Fernando, Busra, & Sulistiani, Heni. (2020). Implementasi Algoritma Multiclass SVM pada Opini Publik Berbahasa Indonesia di Twitter. *Jurnal Tekno Kompak*, *14*(2), 86–91.

Ariyanti, Dyah, & Iswardani, Kurnia. (2020). Teks Mining untuk Klasifikasi Keluhan Masyarakat Menggunakan Algoritma Naive Bayes. *Ikraith-Informatika*, *4*(3), 125–132.

Balan, U. Mahesh, & Mathew, S. K. (2015). Online word of mouth using text mining: A review of literature and future directions. *2015 IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions (WCI)*, 1–6. IEEE.

Darwis, Dedi, Siskawati, Nery, & Abidin, Zaenal. (2021). Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional. *Jurnal Tekno Kompak*, *15*(1), 131–145.

Hussein, Doaa Mohey El Din Mohamed. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences*, *30*(4), 330–338.

Mailani, Okarisma, Nuraeni, Irna, Syakila, Sarah Agnia, & Lazuardi, Jundi. (2022). Bahasa sebagai alat komunikasi dalam kehidupan manusia. *Kampret Journal*, *1*(2), 1–10. https://doi.org/10.35335/kampret.v1i1.8

Nurzahputra, Aldi, & Muslim, Much Aziz. (2016). Analisis sentimen pada opini mahasiswa menggunakan natural language processing. *Seminar Nasional Ilmu Komputer (SNIK 2016)*, 114–118.

Pristiyanti, Ria Ine, Fauzi, Mochammad Ali, & Muflikhah, Laili. (2018). Sentimen Analisis Peringkasan Review Film Menggunakan Metode Information Gain dan K-Nearest Neighbor. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer E-ISSN*, *2548*, 964x.

Rintyarna, Bagus Setya. (2016). Sentiment Analysis pada Movie Review dengan Pendekatan Klasifikasi dalam Algoritma J. 48. *JUSTINDO (Jurnal Sistem Dan Teknologi Informasi Indonesia)*, *1*(2). https://doi.org/10.32528/justindo.v1i2.567

Rintyarna, Bagus Setya. (2022). Joint Distribution pada Weighted Majority Vote (WMV) untuk Peningkatan Kinerja Sentiment Analysis Tersupervisi pada Dataset Twitter. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *9*(5), 1083–1090.

Savitri, Ni Luh Putu Chandra, Rahman, Radya Amirur, Venyutzky, Reyhan, & Rakhmawati, Nur Aini. (2021). Analisis klasifikasi sentimen terhadap sekolah daring pada twitter menggunakan Supervised Machine Learning. *Jurnal Teknik Informatika Dan Sistem Informasi*, *7*(1).

Buniar, Eka, Utsalinah, Dwi Safiroh, & Wahyuningsih, Dian. (2022). Implementasi Scrapping Data Untuk Sentiment Analysis Pengguna Dompet Digital dengan Menggunakan Algoritma Machine Learning. *Jurnal Janitra Informatika Dan Sistem Informasi*, *2*(1), 35–42.