INNOVATIVE: Journal Of Social Science Research
Volume 4 Nomor 3 Tahun 2024 Page 7273-7286
E-ISSN 2807-4238 and P-ISSN 2807-4246
Website: https://j-innovative.org/index.php/Innovative

Eksplorasi Analisis Sentimen pada Rating Film IMDb: Pendekatan Perbandingan menggunakan Decision Tree dan Naive Bayes

Naufal Jakfar Ramadhan[1✉], Veni Agustina Pratama Putri[2], Dimas Ananda Riyadi[3]
(1) Institut Teknologi Telkom Purwokerto
(2) Universitas Terbuka
(3) Universitas Pancasakti Bekasi
Email : 20103108@ittelkom-pwt.ac.id[1✉]

Abstrak

Film merupakan salah satu bentuk hiburan yang sangat populer di dunia saat ini. IMDb hadir sebagai platform terkemuka dalam industri film yang memungkinkan pengguna memberikan rating dan ulasan tentang film yang mereka tonton. Analisis sentimen pada rating film IMDb menjadi penting untuk memahami respon penonton terhadap film-film tersebut. Eksplorasi ini membandingkan performa Decision Tree dan Naive Bayes dalam melakukan analisis sentimen pada rating film IMDb. Decision Tree menggunakan struktur pohon keputusan untuk mengklasifikasikan data, sementara Naive Bayes menggunakan teorema Bayes untuk menghitung probabilitas kelas dari teks. Dengan membandingkan kedua metode ini, eksplorasi ini memberikan wawasan berharga bagi pembuat film, produsen, dan penonton dalam memahami bagaimana film-film tersebut diterima oleh masyarakat secara keseluruhan.
Kata Kunci: *Analisis sentimen, Dataset rating film, Decision Tree, Film, Hiburan, IMDb, Kualitas film, Metode klasifikasi, Naive Bayes, Pembuat film, Penonton, Performa, Pohon keputusan, Produsen, Rating film, Respon penonton, Teorema Bayes, Ulasan pengguna, Wawasan berharga, Pengalaman menonton.*

Abstract

Movies are one of the most popular forms of entertainment worldwide. IMDb serves as a leading platform in the film industry, allowing users to provide ratings and reviews for the movies they watch. Sentiment analysis on IMDb movie ratings becomes crucial in understanding audience responses to these films. This exploration compares the performance of Decision Tree and Naive Bayes in conducting sentiment analysis on IMDb movie ratings. Decision Tree utilizes a decision tree structure to classify data, whereas Naive Bayes uses Bayes' theorem to calculate the probability of text classes. By comparing these two methods, this exploration provides valuable insights for filmmakers, producers, and audiences in comprehending how these films are perceived by society as a whole.

Keywords: *Analysis of sentiment, Audience, Audience response, Bayes' theorem, Classification method, Decision Tree, Decision tree, Entertainment, Film, Film maker, Film quality, Film rating, Film rating dataset, IMDb, Naive Bayes, Performance, Producer, User reviews, Valuable insights, Viewing experience.*

PENDAHULUAN

Film merupakan salah satu bentuk hiburan yang sangat populer di dunia saat ini. Dalam dunia yang penuh dengan keragaman, setiap orang memiliki preferensi dan pendapat mereka sendiri tentang film yang mereka tonton. Beberapa orang mungkin menyukai film aksi yang penuh dengan kegembiraan dan adegan laga, sementara yang lain mungkin lebih tertarik dengan film drama yang menggugah emosi. Dalam keadaan seperti ini, IMDb hadir sebagai salah satu platform terkemuka dalam industri film yang memberikan kesempatan kepada pengguna untuk memberikan rating dan ulasan tentang film yang mereka saksikan (Ramadhan & Ramadhan, 2022). IMDb menjadi tempat di mana penggemar film dari seluruh dunia dapat berbagi pendapat mereka tentang film-film yang telah mereka tonton. Dengan adanya platform ini, pengguna IMDb dapat memberikan rating berdasarkan pengalaman mereka saat menonton film. Mereka juga dapat menulis ulasan yang mendalam tentang apa yang mereka suka atau tidak suka dari film tersebut (Pandunata et al., 2023). Hal ini memberikan kesempatan kepada pengguna lain untuk mendapatkan wawasan tentang film yang ingin mereka tonton sebelum mereka benar-benar menontonnya. Dalam industri film yang terus berkembang, IMDb menjadi sumber informasi yang sangat berharga (Tarımer et al., n.d.).

Namun, dengan banyaknya film yang dirilis setiap tahun, menjadi tantangan bagi para produsen film, pembuat kebijakan, dan penonton untuk memahami bagaimana film tersebut diterima oleh masyarakat. Inilah mengapa analisis sentimen pada rating film IMDb menjadi penting (Çizmeci & Öüdücü, n.d.). Untuk menganalisis sentimen dari ulasan pengguna, kita dapat memperoleh wawasan yang berharga tentang apakah film tersebut mendapatkan respon positif, negatif, atau netral dari penonton(Rizal et al., 2023).

Eksplorasi ini akan menggunakan pendekatan perbandingan antara Decision Tree dan Naive Bayes dalam melakukan analisis sentimen pada rating film IMDb(Topal & Ozsoyoglu, 2016). Decision Tree merupakan algoritma pembelajaran mesin yang menggunakan struktur pohon keputusan untuk mengklasifikasikan data(Pradeep et al., 2020). Dalam eksplorasi ini, kita akan menggunakan pendekatan perbandingan antara Decision Tree dan Naive Bayes untuk melakukan analisis sentimen pada rating film IMDb (Bristi et al., n.d.). Decision Tree adalah salah satu algoritma pembelajaran mesin yang sangat populer. Algoritma ini menggunakan struktur pohon keputusan untuk mengklasifikasikan data. Dengan menggunakan Decision Tree, kita dapat mengidentifikasi pola-pola dalam data rating dan ulasan pengguna untuk memprediksi sentimen yang terkandung di dalamnya. Selain Decision Tree, kita juga akan menggunakan metode Naive Bayes dalam eksplorasi ini (Su & Zhang, n.d.). Naive Bayes adalah metode klasifikasi statistik yang menggunakan teorema Bayes untuk menghitung probabilitas kelas dari suatu teks. Dengan menggunakan Naive Bayes, kita dapat menghitung probabilitas bahwa suatu ulasan pengguna memiliki sentimen positif, negatif, atau netral berdasarkan kata-kata yang terkandung di dalamnya. Dengan membandingkan performa antara Decision Tree dan Naive Bayes, kita akan dapat menentukan metode mana yang lebih efektif dalam melakukan analisis sentimen pada rating film IMDb (Rizal et al., 2023). Hasil dari eksplorasi ini dapat memberikan wawasan yang berharga bagi para pembuat film, produsen, dan penonton dalam memahami bagaimana film-film tersebut diterima oleh masyarakat secara keseluruhan. Sementara itu, Naive Bayes adalah metode klasifikasi statistik yang menggunakan teorema Bayes untuk menghitung probabilitas kelas dari suatu teks. Dalam eksplorasi ini, akan menggunakan dataset rating film IMDb sebagai sumber data (Rish, n.d.). Tujuannya adalah untuk menentukan metode mana yang lebih efektif dalam menganalisis sentimen pada rating film IMDb. Hasil dari eksplorasi ini diharapkan dapat memberikan wawasan yang berharga bagi para produsen film, pembuat kebijakan, dan penonton dalam memahami bagaimana film-film yang mereka produksi atau tonton diterima oleh masyarakat. Dengan demikian, eksplorasi ini dapat berkontribusi dalam meningkatkan kualitas film yang diproduksi serta pengalaman menonton para penonton (Pradeep et al., 2020).

METODE PENELITIAN

Penelitian ini bertujuan untuk mengeksplorasi analisis sentimen pada rating film IMDb melalui pendekatan perbandingan antara metode Decision Tree dan Naive Bayes. Pendekatan ini akan menggali berbagai ulasan dan rating yang diberikan pengguna pada film-film di platform IMDb. Metode Decision Tree akan digunakan untuk membangun

model yang dapat memprediksi sentimen positif atau negatif berdasarkan fitur-fitur yang diambil dari ulasan pengguna. Di sisi lain, metode Naive Bayes akan digunakan sebagai perbandingan untuk mengidentifikasi keunggulan dan kelemahan keduanya dalam analisis sentimen terhadap rating film IMDb. Dengan demikian, penelitian ini diharapkan dapat memberikan pemahaman yang lebih dalam mengenai kinerja dan keefektifan kedua metode tersebut dalam menganalisis sentimen pada platform IMDb.

A. Analisis Sentimen

Analisis sentimen pada rating film IMDb melibatkan pemahaman opini atau sentimen yang terkandung dalam ulasan atau rating yang diberikan oleh pengguna. Metode yang umum digunakan dalam analisis sentimen ini antara lain Decision Tree dan Naive Bayes.

B. Decision Tree

Pohon keputusan merupakan model prediktif yang menggunakan struktur pohon untuk membuat keputusan berdasarkan serangkaian aturan dan pemisahan data berdasarkan fitur yang relevan. Dalam analisis sentimen IMDb, pohon keputusan dapat memisahkan ulasan menjadi kategori positif, negatif, atau netral berdasarkan fitur-fitur tertentu seperti kata-kata kunci, sentimen, atau fitur lainnya.

C. *Naive Bayes*

Naive Bayes adalah metode klasifikasi berdasarkan teorema probabilitas Bayes. Meskipun sederhana, Naive Bayes efektif dalam analisis sentimen karena menghitung probabilitas bahwa suatu ulasan termasuk dalam kategori tertentu berdasarkan kemungkinan kata-kata atau fitur yang muncul dalam ulasan tersebut. Ini dianggap "naif" karena asumsi bahwa fitur-fitur (kata-kata dalam kasus analisis sentimen) adalah independen satu sama lain.

Dalam eksplorasi analisis sentimen menggunakan metode ini, langkah-langkah umum meliputi:

1) Pemrosesan Data: Ulasan dari IMDb harus diproses, termasuk pembersihan teks, penghapusan stop word, tokenisasi, dan vektorisasi (mengubah teks menjadi representasi numerik).
2) Pembagian Data: Data kemudian dibagi menjadi set pelatihan (training set) dan set pengujian (testing set) untuk melatih model dan menguji kinerjanya.
3) Pelatihan Model: Model Decision Tree dan Naive Bayes akan dilatih menggunakan set pelatihan. Dalam pelatihan, model akan belajar untuk mengenali pola-pola yang ada di dalam data latih.

4) Evaluasi Model: Setelah pelatihan, model akan dievaluasi menggunakan set pengujian untuk melihat seberapa baik model tersebut memprediksi sentimen dari ulasan IMDb. Metrik seperti akurasi, presisi, recall, dan F1-score dapat digunakan untuk mengevaluasi kinerja model.
5) Optimasi dan Penyetelan: Kadang-kadang, parameter model dapat disesuaikan atau fitur-fitur tambahan dieksplorasi untuk meningkatkan kinerja model.

D. Analisis Sentimen pada IMDb

Analisis sentimen pada IMDb Ratings mencakup proses mengekstraksi, menginterpretasi, dan memahami sentimen yang terkandung dalam ulasan atau rating yang diberikan oleh pengguna IMDb terhadap film. IMDb adalah platform yang mengumpulkan ulasan, peringkat, dan pendapat pengguna tentang film, dan analisis sentimen pada IMDb Ratings fokus pada mengevaluasi apakah ulasan tersebut bersifat positif, negatif, atau netral terhadap suatu film.

Langkah-langkah dalam Analisis Sentimen IMDb Ratings:

1) Pengumpulan Data

Data yang digunakan untuk analisis sentimen diambil dari ulasan pengguna pada IMDb. Ini bisa berupa teks ulasan, peringkat numerik, atau keduanya. Pengumpulan data ini bisa dilakukan secara otomatis melalui API IMDb atau dengan teknik web scraping.

2) Pemrosesan Teks

Teks ulasan sering kali memerlukan pembersihan dan pemrosesan sebelum analisis dilakukan. Ini melibatkan langkah-langkah seperti menghapus karakter khusus, mengubah teks menjadi huruf kecil, menghilangkan stopwords (kata-kata umum yang tidak memberikan informasi penting), dan melakukan tokenisasi (memisahkan teks menjadi token atau kata-kata individual).

3) Analisis Sentimen

Setelah pemrosesan teks, dilakukan analisis sentimen menggunakan teknik-teknik seperti

a) Lexicon-based Sentiment Analysis: Memanfaatkan kamus kata-kata dengan sentimen yang sudah diketahui (positif, negatif, netral) untuk menilai sentimen dalam teks.
b) Machine Learning (ML) Approaches: Menerapkan model pembelajaran mesin seperti Naive Bayes, Logistic Regression, atau Neural Networks untuk mengklasifikasikan ulasan ke dalam kategori sentimen berbeda.

4) Klasifikasi Sentimen

Setelah analisis, ulasan diklasifikasikan ke dalam kategori sentimen tertentu (positif, negatif, netral) berdasarkan hasil analisis yang dilakukan.

5) Evaluasi dan Interpretasi Hasil

Metrik evaluasi seperti akurasi, presisi, recall, dan F1-score digunakan untuk menilai kinerja model atau teknik yang digunakan dalam analisis sentimen IMDb Ratings. Hasil ini kemudian diinterpretasikan untuk mendapatkan pemahaman yang lebih baik tentang bagaimana ulasan di IMDb merespons suatu film.

E. Evaluasi Model

Evaluasi model dalam konteks analisis sentimen IMDb Ratings atau dalam analisis sentimen secara umum melibatkan langkah-langkah untuk menilai seberapa baik model tersebut mampu memprediksi atau mengklasifikasikan sentimen dari ulasan atau data yang diberikan. Berikut adalah beberapa metrik evaluasi yang umum digunakan:

1) Akurasi (*Accuracy*)

Akurasi mengukur seberapa sering model benar dalam memprediksi sentimen secara keseluruhan dari semua prediksi yang dilakukan.

2) Presisi (*Precision*)

Presisi mengukur seberapa banyak dari prediksi positif yang dilakukan oleh model adalah benar. Dalam konteks analisis sentimen, presisi mengukur seberapa tepat model dalam mengidentifikasi sentimen positif dari ulasan.

3) *Recall* (*Recall/Sensitivity*)

Recall mengukur seberapa banyak dari keseluruhan kelas positif yang dapat diidentifikasi oleh model. Dalam analisis sentimen, ini mengukur seberapa baik model mengenali semua ulasan yang seharusnya diklasifikasikan sebagai positif.

4) F1-Score

F1-Score merupakan perpaduan antara presisi dan recall. Ini berguna ketika ada ketidakseimbangan kelas (imbalance class) dalam data. F1-Score memberikan nilai yang seimbang antara kedua metrik tersebut.

5) 5. *Confusion Matrix*

Matriks kebingungan menampilkan performa model dalam mengklasifikasikan setiap kelas. Ini memungkinkan kita untuk melihat seberapa baik model dapat membedakan antara kelas positif, negatif, dan netral.

6) Kurva ROC (*Receiver Operating Characteristic*)

Kurva ROC mengukur kinerja model untuk berbagai ambang batas (threshold) dalam klasifikasi. Area di bawah kurva ROC (AUC-ROC) memberikan gambaran tentang seberapa baik model membedakan antara kelas positif dan negatif.

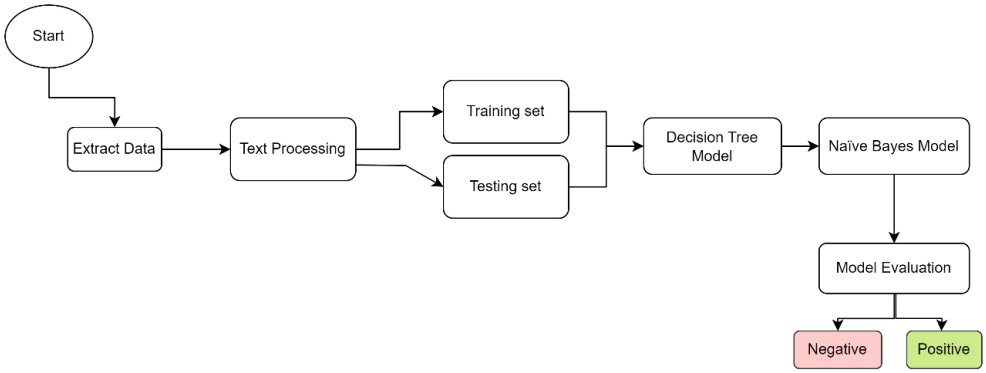
F. Langkah Evaluasi Model

1) Pembagian Data: Data dibagi menjadi set pelatihan (training set) dan set pengujian (testing set).
2) Pelatihan Model: Model dilatih menggunakan data pelatihan dengan teknik-teknik seperti Naive Bayes, Decision Trees, atau algoritma Machine Learning lainnya.
3) Pengujian Model: Model diuji menggunakan data pengujian untuk menghasilkan prediksi sentimen. Hasil prediksi ini kemudian dievaluasi menggunakan metrik-metrik evaluasi yang telah disebutkan di atas.
4) Penyetelan Model (Opsional): Kadang-kadang, parameter model disesuaikan untuk meningkatkan kinerja, terutama dalam algoritma Machine Learning yang memiliki parameter yang dapat diatur.
5) Interpretasi Hasil: Hasil evaluasi digunakan untuk memahami seberapa baik model dalam mengklasifikasikan sentimen. Ini membantu untuk mengevaluasi apakah model dapat digunakan secara efektif dalam memprediksi sentimen ulasan IMDb atau tidak.

Evaluasi model penting untuk memastikan model yang digunakan dapat memberikan prediksi sentimen yang akurat dan dapat diandalkan. Metrik evaluasi memberikan gambaran yang lebih komprehensif tentang kekuatan dan kelemahan model dalam melakukan klasifikasi sentimen.

HASIL DAN PEMBAHASAN

Semua metodologi yang digunakan pada penelitian ini akan dijelaskan pada bagian ini. Konsep metodologi yang digunakan pada gambar 1.0. Metodologi yang akan dilakukan pada penelitian ini yang pertama adalah Extract Data, Text Processing, Training dan Testing set, penggunaan model *Decision Tree* dan *Naïve Bayes*, melakukan evaluasi hasil kerja setiap algortima.



Gambar 1. Diagram Alir

Alur kerja diatas akan diulas pada sub-bab dibawah ini

A. Deskripsi Data

Data yang akan digunakan pada penelitian kali ini seperti yang sudah disebutkan diatas yakni dari IMBd terkait *review* untuk sebuah film. Dataset ini berjenis berkas txt. Txt adalah sebuah berkas teks yang memuat sekumpulan teks sederhana tanpa adanya gambar, video, music, dsb didalamnya. Berkas inilah yang biasa digunakan untuk melakukan Analisa lebih lanjut dalam *text processing.* Berikut adalah tampilan isi dalam dataset yang akan ditelaah pada penelitian saat ini.

Row...	A very, very, very slow-moving, aimless movie about a ... String	0 Number (integer)
Row0	Not sure who was more lost - the flat characters or the audience, ne...	0
Row1	Attempting artiness with black & white and clever camera angles, th...	0
Row2	Very little music or anything to speak of.	0
Row3	The best scene in the movie was when Gerardo is trying to find a so...	1
Row4	The rest of the movie lacks art, charm, meaning... If it's about empti...	0
Row5	Wasted two hours.	0
Row6	Saw the movie today and thought it was a good effort, good messag...	1
Row7	A bit predictable.	0
Row8	Loved the casting of Jimmy Buffet as the science teacher.	1
Row9	And those baby owls were adorable.	1
Row10	The movie showed a lot of Florida at it's best, made it look very appe...	1
Row11	The Songs Were The Best And The Muppets Were So Hilarious.	1
Row12	It Was So Cool.	1

Gambar 2. Dataset : imdb_labelled.txt

Dalam dataset tersebut terdapat 2 kolom dengan memiliki 2 jenis tipe data yang berbeda. Untuk kolom pertama sebelah kiri yakni *A very, very, very slow-moving, aimless movie about a distressed, drifting young man* dengan jenis tipe data string. String adalah tipe data berupa teks atau huruf. Dari karakteristik pada setiap nilai yang ada pada *imdb_labelled.txt* dapat dilihat bahwa kolom tersebut adalah sekumpulan komentar dari para pengguna terkait sebuah film yang mereka tonton, dan terakhir adalah kolom 0, terdapat dua jenis nilai yang ada pada *variable* ini berdasarkan asal dari dataset ini menjelaskan bahwa 0 yang memiliki arti *negative* dan 1 yang bermakna *positive.*

B. Pemrosesan Data

Dalam *natural language processing* ada tiga langkah penting yang membuka jalan bagi analisis dan pemodelan yang sukses yaitu : *Enrichment, Preprocessing, Transformation*. Tahapan – tahapan yang telah disebutkan adalah tahapan yang dilakukan pada penelitian ini:

1. Enrichment

   Enrichment atau pengayaan adalah Teknik yang dapat meningkatkan konten informasi dari data teks, menjadikannya lebih berharga untuk tugas-tugas NLP. Proses ini terutama melibatkan tambahan pengetahuan dan fitur terbaru yang tidak secara eksplisit ada dalam teks mentah. Pada penelitian kali ini menggunakan Part Of Speech untuk menetapkan kategori tata bahasa (misalnya, kata benda, kata kerja, kata sifat) untuk setiap kata dalam teks.

2. Preprocessing

   Preprocessing adalah tahap membersihkan dan menyiapkan data teks untuk analisis lebih lanjut. Hal ini bertujuan untuk menghilangkan noise, ketidakkonsistenan, dan informasi yang tidak relevan yang dapat menghambat tugas-tugas hilir. Langkah-langkah prapemrosesan yang dilakukan meliputi: 1)menghapus tanda baca, 2)menghapus nomor, 3)menghapus kata yang kurang dari 3 huruf, 4) menghapus kata yang tidak memiliki arti seperti "the", "a", "end", dan sebagainya. 5). Menjadikan semua huruf kecil, 6). Mengurangi kata-kata ke *Stem,*7)menghapus semua kata selain kata sifat, kata keterangan, dan kata benda.

3. *Transformation and Frequencies*

   Transformation atau Transformasi melibatkan pemetaan data tekstual yang cocok untuk algoritme pembelajaran mesin. Sementara Frequencies adalah tahapan yang melibatkan penghitungan kemunculan kata-kata individual dalam data teks. Hal ini membantu mengidentifikasi kata kunci, pola, dan topik yang dominan. Kali ini peneliti melakukan : 1). Membuat *Bag of words* 2). Menghitung frekuensi istilah relative, 3). Mendapatkan representasi vektor dari setiap dokumen.

C. Analisis Sentimen pada Decision Tree

Algoritma pertama yang digunakan adalah *Decision Tree* atau Pohon Keputusan, peneliti melakukan partisi data yang sudah bersih menjadi 2 bagian yakni sebanyak 80% untuk *training* atau bahan ajar algoritma, dan 20% untuk *testing* atau pengujian kemampuan model. Disa dilihat pada Gambar 3 bahwa hasil dari belajar mesin pada algoritma *Decision Tree* :

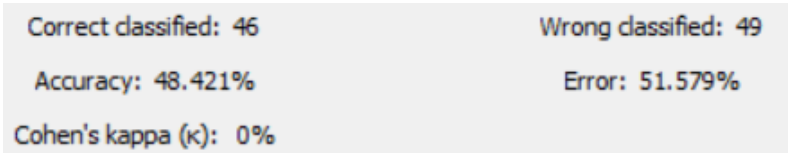
Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (ð□□¹)	Correctly Classified	Incorrectly Classified
71.58%	28.42%	0.423	68	27

Gambar 3 : Decision Tree's score

D. Analisis Sentimen pada Naïve Bayes

Algoritma kedua yang digunakan adalah *Naïve Bayes*, sama halnya seperti pada Algoritma *Decision Tree* peneliti membagi data menjadi 2 bagian yakni 80% sebagai *data training* dan sisanya yakni 20% sebagai *data testing* dan mendapatkan hasil sebagai berikut:



Gambar 4. Result score of Naïve Bayes algorithm

E. Perbandingan

Dari hasil penggunaan 2 algoritma yang berbeda ini bisa dilihat ada perbedaan yang sangat signifikan pada kedua algoritma, nilai akurasi dari *decision tree* adalah 71.58%, *cohen's kappa :* 0.423, *overall error :*28.42%, *correctly classified :* 68, *incorrectly classified :* 27. Sementara nilai akurasi yang dihasilkan algoritma *naïve bayes* yakni 48.421%, *correct classified :* 46, *wrong classified :* 49, *error :* 51.579%, *cohen's kappa(k) :* 0%

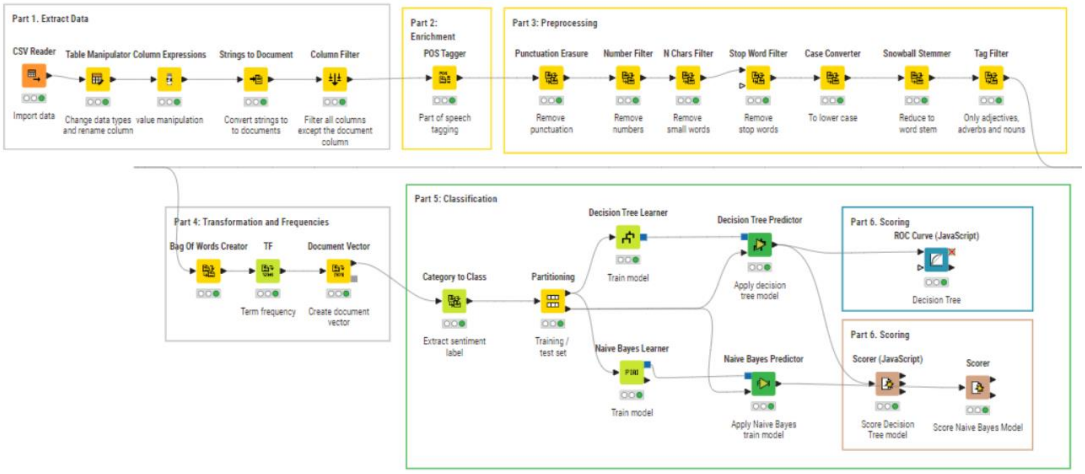
Tabel 1. Perbandingan hasil akurasi dari pemodelan Decision Tree dan Naive Bayes

Algoritma	Correct Classified	Incorrect Classified	Accuracy	Cohen's Kappa (K)	Error
Decision Tree	68	27	71.58%	0.423	28.42%
Naive Bayes	46	49	48.421%	0%	51.579%

F. Pembahasan

Metrik kinerja dari kedua pengklasifikasi menyatakan kesimpulan bahwa pengklasifikasi *decision tree* memberikan hasil terbaik dalam hal *accuracy, cohen's kappa. Correctly classified, overall error* untuk set data IMDb.

Hasil pada penelitian Eksplorasi Analisis Sentimen pada Rating Film IMDb: Pendekatan Perbandingan menggunakan Decision Tree dan Naive Bayes adalah sebagai berikut pada gambar 5:



Gambar 5. Workflow

Pada gambar 5 diatas, terdapat 6 tahapan yang telah dilakukan yakni sebagai berikut :

1. Extract Data

   Terdapat CSV Reader Node yang berguna untuk memasukkan data dari penyimpanan komputer ke dalam platform KNIME, karena data masih bersifat kotor dan belum dapat memenuhi kriteria untuk dilakukan step selanjutnya, maka dari itu perlu dilakukan ekstraksi berupa Table Manipulator sebagai perubah nama kolom yang semula berbentuk "A very, very, very slow-moving, aimless,.." menjadi "review" dilakukan perubahan pada nama kolom agar mempermudah mengenali kolom dalam proses analisa selanjutnya dan mengonversi tipe data pada kolom "category" yang semula *numeric* menjadi *string*. Hal ini dilakukan karena tipe data pada kolom tersebut akan mendukung proses perubahan label pada nilai data yang terdapat pada kolom "category" mulanya adalah "0" menjadi "NEG", dan "1" dirubah menjadi label "POS" menggunakan Column Expressions. Tahap ke empat merubah tipe data pada kolom "review" adalah *String*, dalam analisa sentimen tipe data yang memenuhi syarat adalh tipe data dokumen teks, sehingga dilakukan pengubahan menggunakna String to Document Node, hingga sampailah ke penghujung fase pertama ini yakni melakukan pemfilteran terhadap data yang menjadi titik fokus penilitan selanjutnya yakni hanya menjadikan kolom "Document".

2. Enrichment

   Tahap Enrichment kita gunakan untuk melakukan pengelompokkan (tag) pada setiap kata apakah pada kata tersebut termasuk kedalam kata benda, kata sifat, kata bantu, dan seterusnya menggunakan metode POS Tagging.

3. Prepocessing

   Fase Prepocessing menjadi tahapan yang paling panjang yang mempunyai 7

bagian yakni pertama melakukan penghapusan tanda baca, nomor-nomor yang tidak diperlukan, kata-kata yang hanya memiliki sedikit huruf seperti sebagai contoh : "you", "and", "for", dan lain sebagainya. Selanjutnya adalah menghapus kata-kata yang tidak memiliki arti dalam setiap kalimat, menyamaratakan huruf besar yang ada pada setiap kalimat menjadi huruf kecil semua, lalu meriduksi *stemming* kata mecin bisa menerjemahkan kata dengan lebih baik, dan yang terakhir adalah setelah melakukan pengelompokkan (tag) tahap terakhir pada fase prepocessing adalah melakukan pemfilteran dengan mengambil kata yang berada pada kelompok kata sifat, kata benda, dan kata bantu.

4. Transformation and Frequencies

Fase ini ada tiga tahap dimana tahapan yang pertama adalah menggabungkan file Dokumen yang sudah dibuat pada tahap Extract Data dan file yang ada pada tahap Perpocessing. Dua hasil ekstraksi tersebut digabung menjadi 1 data dalam bentuk kolom bertipe data Term menggunakan Bag of Words Creator, lalu setelah pada data Term di hitung bobot nilai per-row nya menggunakan TF Node dengan rumus sebagai berikut :

$$TF - IDF\ (t,d) = TF(t,d) \times IDF(t) \quad (1)$$

$$TF(t,d) = \frac{n_t}{n} \quad (2)$$

$$IDF(t) = \frac{N_d}{N} \quad (3)$$

Nilai yang memiliki bobot tertinggi akan di pilih menjadi nilai yang memiliki makna penting pada setiap katanya. Lalu lanjut dengan merubahnya kedalam nilai vektor, yang bertujuan untuk mempermudah mesin mempelajari data sehingga tidak memakan waktu banyak dalam menganalisa dan akan memberikan hasil yang lebih akurat, serta sebagai alat pencocokan apakah dokumen yang satu dengan dokumen yang lainnya dalam data memiliki kesamaan atau tidak.

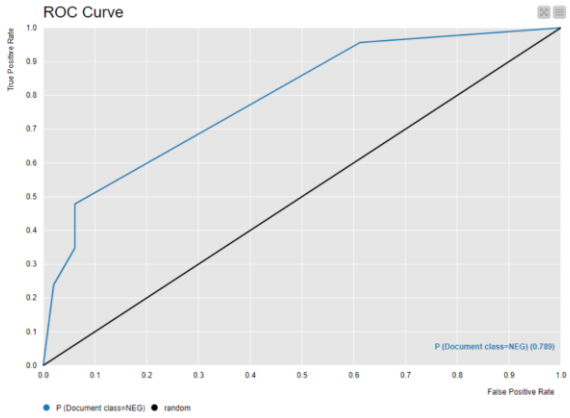
5. Classification

Mengekstrak kolom "category" ke dalam sebuah label sehingga dapat dijadikan titik acuan untuk pembelaran mesin dalam melakukan klasifikasi menggunakan Category to Class node, menggunakan Partitioning node agar dapat membagi data menjadi 2 bagian yakni data pelatihan (training) dan data pengujian (testing), pada babak ini bagian pelatihan hanya diambil 80% sementara data latihan menjadi 20%, yang mengacu aturan 80:20 pada *The Pareto Principle* menyatakan bahwa sekitar 80% dampak berasal dari 20% penyebab. Ada 2 jenis algoritma

pembelajaran mesin yang dilakukan pada penelitian ini yaitu Decision Tree dan Naive Bayes. Pada gambar diatas terdapat Decision Tree Learner Node sebagai algoritma yang bertugas untuk mempelajari data yang sudah bersih dan dilanjut dengan Decision Tree Predictor Node Sebagai alat untuk menguji kecerdasan model. Tahapan yang sama juga berlaku pada pemodelan Naive Bayes.

6. Scoring

Dan yang terakhir menuju kedalam tahap Scoring, dimana melakukan perbandingkan pada 2 jenis algoritma yang telah digunakan pada penilitan kali ini yaitu Decision Tree dan Naive Bayes dengan hasil skor tertinggi adalah pada algoritma Decision Tree. Pada gambar 6 terlihat tingkat akurasi pada ROC Curve hampir membentuk sudut siku-siku, dengan sudut *True Positive Rate* (TPR) berapa di bawah titik 0.5 yang artinya nilai t adalah 0,65 dari sekali 0-1.



Gambar 6 Visualization of Decision Tree

SIMPULAN

Hasil dari penelitian bahwa, telah melakukan analisis sentimen terhadap dataset IMDb yang berisi ulasan tentang sebuah film. Tujuan utama adalah mengklasifikasikan kalimat-kalimat dalam ulasan tersebut menjadi positif atau negatif. Untuk mencapai tujuan ini, menggunakan dua algoritma klasifikasi, yaitu *Decision Tree* dan *Naive Bayes*. Hasil analisis menunjukkan bahwa *Decision Tree* muncul sebagai algoritma yang memberikan akurasi terbaik untuk dataset di penelitian ini. Tentu menandakan kemampuan *Decision Tree* untuk memahami dan mengklasifikasikan sentimen dengan baik dalam konteks ulasan film IMDb. Namun, temuan menarik lainnya muncul dalam proses ini adalah meskipun *Decision Tree* memberikan performa yang sangat baik, kami meyakini bahwa potensi peningkatan skor terbaik dapat dicapai melalui penambahan data. Pengumpulan lebih banyak data dapat memberikan pemahaman yang lebih mendalam terhadap variasi sentimen dalam

ulasan, sehingga meningkatkan kemampuan model untuk menggeneralisasi dan mengklasifikasikan dengan lebih tepat. Tantangan selanjutnya dalam penelitian ini adalah untuk terus memperluas dan memperkaya *dataset*. Untuk saat ini berencana agar melibatkan lebih banyak ulasan, termasuk dari berbagai genre dan periode waktu, agar model dapat menjadi lebih tangkas dan dapat diandalkan dalam mengklasifikasikan sentimen pada ulasan film IMDb.

DAFTAR PUSTAKA

Bristi, W. R., Zaman, Z., & Sultana, N. (n.d.). *Predicting IMDb Rating of Movies by Machine Learning Techniques*. https://www.rottentomatoes.com

Çizmeci, B., & Öüdücü, G. (n.d.). *Predicting IMDb Ratings of Pre-release Movies with Factorization Machines Using Social Media*.

Pandunata, P., Nurdiansyah, Y., & Alfina, F. D. (2023). Aspect-Based Sentiment Analysis of Avatar 2 Movie Reviews on IMDb Using Support Vector Machine. *E3S Web of Conferences*, *448*, 02041. https://doi.org/10.1051/e3sconf/202344802041

Pradeep, K., Tinturosmin, C. R., Durom, S. S., & Anisha, G. S. (2020). Decision Tree Algorithms for Accurate Prediction of Movie Rating. *Proceedings of the 4th International Conference on Computing Methodologies and Communication, ICCMC 2020*, 853–858. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-000158

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

Rish, I. (n.d.). *An empirical study of the naive Bayes classifier*.

Rizal, C., Kifta, D. A., Nasution, R. H., Rengganis, A., & Watrianthos, R. (2023). *Opinion classification for IMDb review based using naive bayes method*. 030025. https://doi.org/10.1063/5.0171628

Su, J., & Zhang, H. (n.d.). *A Fast Decision Tree Learning Algorithm Introduction and Related Work*. www.aaai.org

Tarımer, İ., Çoban, A., & Kocaman, A. E. (n.d.). *Sentiment Analysis on IMDB Movie Comments and Twitter Data by Machine Learning and Vector Space Techniques*.

Topal, K., & Ozsoyoglu, G. (2016). Movie Review Analysis: Emotion Analysis of IMDb Movie Reviews. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.