e-ISSN: 2579-5570
p-ISSN: 2579-4116

Text Mining Untuk Analisis Sentimen Review Film Menggunakan Algoritma Naïve Bayes

Muhammad Haidar Rifki[1], Yustina Retno Wahyu Utami[2], Paulus Harsadi[3]
[1)] Program Studi Sistem Informasi, STMIK Sinar Nusantara
[2,3)] Program Studi Informatika, STMIK Sinar Nusantara
[1-3]Jl. KH. Samanhudi no. 84-86, Laweyan, surakarta
Email: 16400095.muhammad@sinus.ac.id[1] ; yustina_retno@sinus.ac.id[2] ; paulusharsadi@sinus.ac.id[3]

Abstrak - Ketersediaan big data mengarahkan penggunaan teknologi informasi untuk menganalisis data. Salah satu pemanfaatan teknologi informasi untuk mengolah data dalam jumlah besar adalah text mining. Film adalah sesuatu yang dapat Anda tonton di waktu senggang. Saat ini banyak sekali film yang bisa disaksikan melalui internet atau bioskop. Film yang ditonton di internet terkadang dikenakan biaya agar calon penonton membaca komentar dari pengguna yang telah menonton film tersebut. Bisnis film dan reviewnya tidak dapat dipisahkan. Situs ulasan film menjadi sumber ulasan kredibel yang diposting di forum publik. Komentar di website tersebut banyak dan beragam, dan komentar dapat dilihat berdasarkan judul filmnya. Hal ini membuat pengguna kesulitan menganalisis dan menyimpulkan berbagai komentar pengguna lain. Penelitian ini bertujuan untuk menganalisis sentimen opini dari beberapa komentar dan akan mengklasifikasikannya menggunakan metode naïve Bayes classifier. Analisis sentimen adalah proses klasifikasi untuk memahami opini, interaksi, dan emosi suatu dokumen atau teks. Teknik TF-IDF diterapkan untuk mengubah dokumen menjadi beberapa kata dan pengklasifikasi Naïve Bayes digunakan untuk menyelesaikan masalah prediksi kelas jamak. Hasil pengujian dengan konfusi matriks diperoleh akurasi sebesar 87,2%, recall sebesar 92,4%, presisi sebesar 80,9%, dan tingkat error sebesar 12,8%.

Kata kunci: Klasifikasi, *IMDb, Naïve Bayes,* Sentimen Analisis

Abstract-The availability of big data directs the use of information technology to analyze data. One use of information technology to process large amounts of data is text mining. Movies are something you can watch in your leisure time. Currently, there are lots of movies that can be watched via the internet or cinema. Movies watched on the internet are sometimes charged so that potential viewers will read comments from users who have watched the film. The film business and its reviews are inseparable. Movie review sites are becoming a source of credible reviews posted in public forums. The comments are many and varied on the website, and comments can be viewed based on the movie title. This makes it difficult for users to analyze and conclude various comments from other users. This research aims to analyze opinion sentiment from several comments and will classify them using the naïve Bayes classifier method. Sentiment analysis is a classification process for understanding the opinions, interactions, and emotions of a document or text. The TF-IDF technique is applied to transform documents into several words and the Naïve Bayes classifier is used to solve multi-class prediction problems. The test results with the confusion matrix obtained an accuracy of 87.2%, recall of 92.4%, precision of 80.9%, and an error rate of 12.8%.

Keywords: Classification, IMDb, Naïve Bayes, Sentiment Analysis

I. PENDAHULUAN

IMDB adalah situs web yang biasa digunakan untuk melihat diskripsi dari sebuah film seperti aktor/aktris, sinopsis film, rating, serta komentar/*review film*. *Review film* dari situs ini menghasilkan respon positif atau negatif dari para masyarakat yang sudah menonton.

Selama ini produsen film komersil dalam memproduksi film lebih melihat fenomena lingkungan atau mengangkat cerita film dari kisah nyata atau novel popular. Produsen film masih belum menggunakan review oleh penonton untuk menentukan tema, gendre atau cerita film apa yang disenangi oleh penonton. Padahal penonton merupakan konsumen utama sebuah film.

Berdasarkan masalah diatas maka dalam penelitian ini melakukan Analisa sentiment film sehingga bisa digunakan oleh produsen film dalam membuat film yang disukai dipasar.

Respon di IMDB nantinya akan diolah menggunakan text mining untuk menghasilkan sentimen negatif atau positif dari masyarakat sehingga dapat digunakan untuk menilai sebuah film. *Text mining* atau pemanbangan teks adalah proses mengubah teks tidak terstruktur menjadi format terstruktur untuk mengidentifikasi pola bermakna. *Text mining* digunakan untuk menganalisis banyak koleksi materi tekstual untuk menangkap konsep utama, tren, dan pola tersembunyi.

Dalam tahapan *text mining* terdapat proses *preprocessing* yang nanti akan digunakan untuk membuang elemen yang tidak dibutuhkan pada proses penentuan sentimen. Untuk mendapatkan intisari dari data yang diambil pada tahap *preprocessing* terdiri dari beberapa langkah yaitu cleansing, tokenizing, stopword removal, dan stemming. Kemudian data yang sudah melalui tahap *preprocessing* akan diklasifikasikan menggunakan naïve bayes classifier dan seleksi fitur dengan *term frequency*.

Tahap *preprocessing* diterapkan pada data review atau komentar dari situs web *review film*. Selanjutnya fitur yang dihasilkan dari *term frequency* dan *inverse document frequency* digunakan dalam klasifikasi menggunakan *naïve bayes classifier*. Hasil klasifikasi akan menjadi suatu rekomendasi terhadap suatu film tertentu dengan penekanan sentimen pengguna pada situs web *review film*.

Sentimen yang didapatkan melalui situs web review film, dilakukan proses pengolahan menggunakan preprocessing dan naïve bayes classifier sehingga terbentuk perbandingan sentimen yang dapat dijadikan rekomendasi terhadap suatu film. Berdasarkan permasalahan tersebut, diperlukan suatu

e-ISSN: 2579-5570
p-ISSN: 2579-4116

analisis untuk mengetahui akurasi yang terbaik dari penggunaan klasifikasi *naïve bayes* dengan *term frequency* pada suatu sistem rekomendasi.

Algoritma *naïve bayes* dipakai dalam penelitian ini karena cukup efektif dalam proses *text mining* berikut beberapa penelitian menggunakan algoritma naïve bayes.

Analisis sentimen masyarakat menggunakan metode SVM dan *Naïve Bayes Classifier* (NBC) untuk permasalahan kelistrikan di Kota Ambon [1]. Data berupa tweet yang diambil dari Twitter dalam batas waktu tertentu. Hasil analisis ini menunjukkan bahwa terdapat tidak lebih dari 50% sentimen negatif terhadap masalah kelistrikan. Hal ini karena pengaruh adanya pemberian informasi ke masyarakat secara real time mengenai masalah kelistrikan. Beberapa penelitian lain tentang sentimen yang menggunakan algoritma KNN dan Naives Bayes [2][3] dan menggunakan Support Vector Machine (SVM)[4][5][6] .

Tahapan penelitian yang dilakukan merujuk pada penelitian di atas dengan sumber data adalah review di IMDb. Situs IMDb dipilih karena memiliki pengguna aktif paling besar di antara situs web review film.

II. TINJAUAN PUSTAKA

A. *Analisis Sentimen*

Analisis sentimen merupakan bidang penelitian yang meneliti ekstraksi sentimen dalam sebuah kalimat berdasarkan isi. Bidang yang berkaitan dengan penelitian ini adalah *Data Mining, Natural Language Processing* (NLP) dan *Machine Learning* [1], sering disebut juga dengan penambangan opini *(opinion mining)*. Penambangan opini merupakan teknik analisis teks yang menggunakan pemrosesan bahasa alami untuk secara otomatis mengidentifikasi dan mengekstrak sentimen atau opini dari dalam teks (positif, negatif, netral, dll.).

B. *Text Mining*

Text mining adalah adalah proses menemukan informasi dengan mengeksplorasi dan menganalisa data besar yang tidak terstruktur. Proses *text mining* mengasosiasikan satu bagian *text* dengan yang lainnya berdasarkan aturan-aturan tertentu. Hasil yang diharapkan adalah informasi atau pengetahuan baru [7].

C. *Term Frequency*

Term Frequency diperlukan untuk mengetahui suatu pola atau suatu bentuk term untuk membantu proses klasifikasi dalam suatu dokumen. *Term Frequency* akan mendata sejumlah kata yang terdapat dalam dokumen yang beragam dengan hasil yang diharap mendapatkan term yang berbeda [1].

a. Term Frequency (TF) Factor

Faktor *Term Frequency* (tf) diperoleh dengan menghitung frekuensi kemunculan sebuah kata dalam dokumen. Frekuensi kemunculan *term* dalam dokumen akan mempengaruhi bobot term tersebut. Nilai *term frequency* ini digunakan untuk memberi bobot suatu kata. Semakin nilai tf dalam dokumen, akan besar pula bobotnya.

b. Inverse Document Frequency (IDF) Factor

Faktor *Inverse Document Frequency* (IDF) menggambarkan, seberapa umum atau jarang suatu *term* pada suatu kumpulan dokumen. Frekuensi kemunculan *term* yang tinggi di berbagai dokumen, dianggap sebagai *term* umum (*common term)* dan tidak penting. Sebaliknya, frekuensi kemunculan *term* yang rendah dianggap sebagai kata yang lebih penting (*uncommon terms)*. Nilai IDF yang semakin dekat ke 0, menunjukkan semakin umum suatu kata.

D. Naïve Bayes Classifier

Teorema keputusan *Bayes* adalah pendekatan statistik yang mendasar dalam pengenalan pola *(pattern recoginition)*. *Naïve Bayes classifier* adalah algoritma machine learning terawasi yang digunakan untuk mengklasifikasikan sesuatu seperti klasifikasi teks. Dalam pengklasifikasiannya, algoritma ini menggunakan probabilitas atau kemungkinan sesuai dengan teorema Bayes [2].

E. Laplacian Smoothing

Laplacian smoothing adalah Teknik penghalusan untuk menangani permasalahan probabilitas nol di *Naïve Bayes*. Prinsip pengklasifikasian naïve bayes berdasarkan probabilitas berdasarkan kemunculan fitur data pelatihan. Jika suatu fitur tidak muncul di data pelatihan, maka probabilitasnya menjadi nol. Hal ini menyebabkan akan mengalikan probabilitas dengan nol sehingga menghasilkan probabilitas nol secara keseluruhan. Penerapan *laplacian smoothing* akan mencegah probabilitas nol dan memastikan bahwa fitur yang tidak terlihat sekalipun memiliki kemungkinan yang bukan nol [3].

F. *Pengujian*

Pengujian pada penelitian ini untuk mengukur keberhasilan fungsionalitas sistem dan performa dari metode yang dilakukan. Pengujian fungsional sistem menggunakan metode menggunakan blackbox testing edankan pengujian performa menggunakan confusion matrix.

a. Black Box Testing

Blackbox testing adalah pengujian yang didasarkan pada fungsionalitas atau eksekusi terhadap unit/modul dari sistem. Hasilnya apakah sesuai dengan yang diharapkan tanpa melihat ke dalam struktur kode sebuah sistem aplikasi.

b. Confusion Matrix

Pengujian *confussion matrix* digunakan untuk menguji performa metode yang digunakan. Terdapat beberapa faktor yang dapat dilihat dari pengujian ini, yaitu *accuration, recall, precision* dan *error rate*.

e-ISSN: 2579-5570
p-ISSN: 2579-4116

G. Penelitian Terdahulu

Beberapa penelitian yang memiliki kesamaan tema ataupun metode dan menjadi rujukan dari penelitian ini diantaranya:

a. Menurut [5] dan [6], Algoritma *Naive Bayes* dapat digunakan untuk memperkirakan waktu studi mahasiswa. Berdasarkan data kelulusan sebelumnya, setiap mahasiswa yang telah menempuh kuliah hingga minimal semester IV dapat diprediksi lama masa studinya. Teknik yang digunakan adalah teknik data mining dengan pengklasifikasi naïve bayes. Lama masa studi mahasiswa diperkirakan berdasarkan faktor-faktor mengenai latar belakang sekolah sebelumnya, data akademik mahasiswa dan tiap pribadi mahasiswa saat di perguruan tinggi. Hasil klasifikasi memperlihatkan bahwa tingkat kesalahan berkisar antara 20% sampai 34% yang mungkin dipengaruhi oleh jumlah data latih maupun data uji serta tingkat konsisten data yang digunakan [5][11]. Semakin banyak data training yang digunakan maka tingkat *Recall, Precision* dan *accuration* akan semakin baik [6].
b. Studi mengenai pengelompokan keanekaragaman dokumen teks yang jumlahnya sangat besar dan terus menumpuk bertujuan untuk memudahkan pencari informasi dalam mendapatkan informasi yang dibutuhkan [7]. Hasil penelitian ini berupa aplikasi pengelompokan dokumen teks dengan menerapkan metode *winnowing* untuk pemilihan fitur. Penamaan file dokumen terkadang tidak sama dengan isi dokumen, menyebabkan pencarian dokumen dengan menggunakan nama file seringkali mendapatkan hasil yang tidak relevan. Dokumen diberi label atau dikelompokkan berdasarkan isi karakter yang ada di dalamnya akan mempermudah dalam proses pencarian dokumen. Konfigurasi *winnowing* yang tepat akan menghasilkan fitur dokumen terbaik untuk pengelompokan dokumen [7].
c. Studi tentang klasifikasi nasabah dari suatu perusahaan asuransi dilakukan untuk melihat nasabah yang lancar, kurang lancar, atau tidak lancar dalam membayar premi [8]. Hasilnya adalah sebuah sistem pengklasifikasian kelompok nasabah dalam membayar iuran premi asuransi yang termasuk kelompok lancar, kelompok kurang lancar dan kelompok tidak lancar dengan menggunakan algoritma *Naïve Bayes*. Dalam penelitian ini, digunakan variabel jenis pekerjaan, besarnya penghasilan per tahun, masa pembayaran asuransi, dan cara pembayaran asuransi serta jenis kelamin, usia, dan status. Adanya sistem ini memudahkan pengambil keputusan apakah menerima atau menolak calon nasabah [8].
d. Penelitian tentang bagaimana membuat aplikasi untuk mengenali emosi dari kalimat berbahasa Indonesia telah dilakukan [9]. Aplikasi yang dihasilkan dikembangkan menggunakan Matlab berbasis *Graphical User Interface* (GUI). Berdasarkan pengujian, diperoleh akurasi sebesar 88,2% [9].

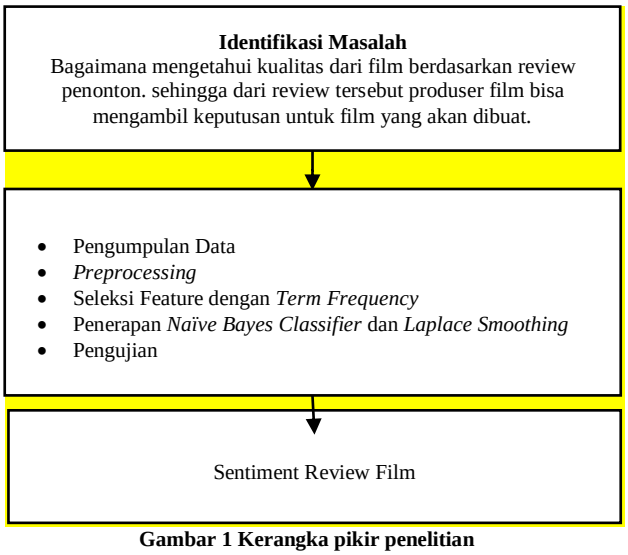
Pada Tabel 1 diperlihatkan mengenai perbandingan perbedaan dan persamaan dengan penelitian terdahulu.

Tabel 1 Perbandingan penelitian

No	Jurnal atau Artikel	Persamaan	Perbedaan
1.	Algoritma Naive Bayes untuk Mencari Perkiraan Waktu Studi Mahasiswa [11]	- Mengguna kan metode Naïve Bayes.	- Perhitungan dilakukan menggunakan membuat fungsi query di mysql. - Pengumpulan data menggunakan excel
2.	Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier [12]	- Mengguna kan metode Naïve Bayes.	- Menggunakan metode pengujian confusion matrix - Menggunakan aplikasi berbasis desktop.
3.	Penerapan Metode Winnowing Fingerprint dan Naive Bayes untuk Pengelompokan Dokumen [13]	- Mengguna kan metode Naïve Bayes.	- Penelitian menggunakan metode winnowing fingerprint - Aplikasi menggunakan web base - Penelitian ini tidak menerapkan filtering pada proses text preprocessing untuk menghilangkan kata penghubung
4.	Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi [16]	- Mengguna kan metode Naïve Bayes.	- Perancangan sebuah sistem aplikasi - Aplikasi berbasis desktop
5.	Pemanfaatan Naïve Bayes Untuk Merespon Emosi Dari Kalimat Berbahasa Indonesia [17]	- Mengguna kan metode Naïve Bayes.	- Penelitian menggunakan aplikasi matlab - Data dokumen yang digunakan berbahasa indonesia

III. METODE PENELITIAN

Pada penelitian ini menggunakan beberapa metode dalam pengumpulan data serta menganalisa data untuk mengambil hasil dan kesimpulan.



Gambar 1 Kerangka pikir penelitian

e-ISSN: 2579-5570
p-ISSN: 2579-4116

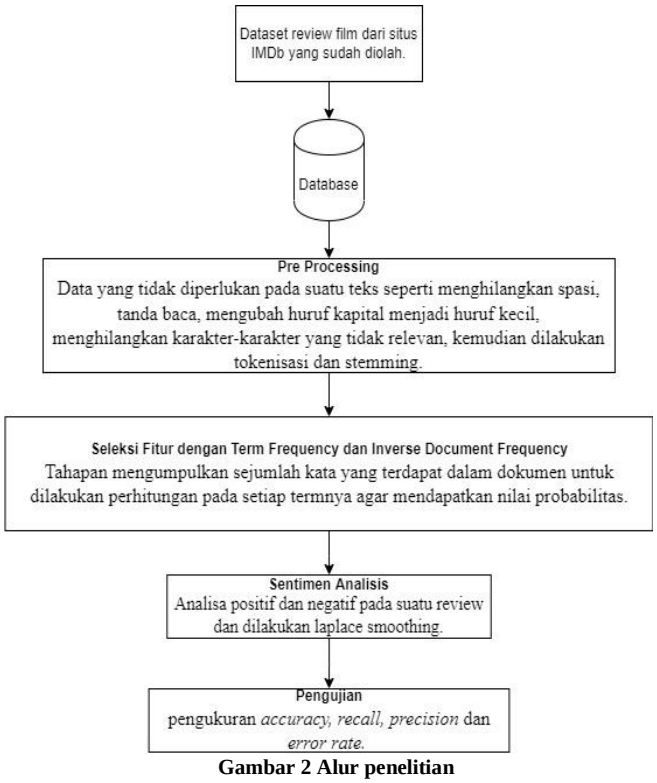
A. Metode Pengumpulan Data

Metode pengumpulan data yang lakukan dalam penelitian ini diantaranya:

a. Metode Studi Literatur
   Studi literatur dilakukan dengan mencari informasi dari buku-buku, jurnal, maupun artikel. Data-data yang diperoleh dalam metode ini berupa data yang terkait dengan sentimen analisis di situs web IMDb dan juga tentang metode yang digunakan, yakni menggunakan metode *Naïve Bayes Classifier*.
b. Metode Observasi
   Metode pengumpulan data sentimen yang didapat menggunakan dataset yang sudah dilakukan pelabelan pada website *http://ai.stanford.edu/~amaas/data/sentiment/*. Data yang sudah diolah tersebut berasal dari situs web review film IMDb. Jumlah data yang sudah diolah sejumlah 5.000 data yang terbagi menjadi dua kelas yaitu Data sentimen positif sebanyak 2500 dan data sentimen negatif sebanyak 2500. Dataset yang digunakan mempunyai dua kelas yaitu review dan sentimen.

B. Teknik Pengolahan Data

Alur yang digunakan untuk melakukan pengolahan data penelitian ditunjukan pada Gambar 2 dibawah ini.



Gambar 2 Alur penelitian

Proses awal yang dilakukan adalah dengan mendapatkan dataset atau informasi review dari situs web IMDb yang sudah diolah oleh penelitian sebelumnya dari Stanford University. Setelah mendapatkan data review, lalu disimpan ke database. Dimana didalam beberapa tabel database sudah disiapkan beberapa *field* yang ditunjukan pada Tabel 2, 3 dan 4 dibawah ini.

Tabel 2 Tabel Datasets

Field	Deskripsi	Tipe Data
Id	Merupakan nomor unik untuk setiap data yang disimpan	Text
Review	Konten review yang ditulis oleh pengguna	Text
Sentiment	Kategori suatu review yang berupa positif atau negatif	Text

Tabel 3 Tabel Text Processing

Field	Deskripsi	Tipe Data
Id	Merupakan nomor unik untuk setiap data yang disimpan	Text
Review	Teks sentimen sebelum dilakukan preprocessing	Text
textProcessed	Teks sentimen sesudah dilakukan preprocessing	Text
Sentiment	Kategori suatu review yang berupa positif atau negatif	Text

Kemudian untuk field pada tabel klasifikasi sama dengan tabel *text processing*. Tetapi value nya ada beberapa yang beda karena sudah dilakukan proses *text processing* sebelumnya.

Tabel 4 Tabel Klasifikasi

Field	Deskripsi	Tipe Data
Id	Merupakan nomor unik untuk setiap data yang disimpan	Text
Review	Teks review yang sudah dilakukan preprocessing	Text
textProcessed	Teks sentimen sesudah dilakukan preprocessing	Text
Sentiment	Kategori suatu review yang berupa positif atau negatif	Text

Kemudian adalah melakukan tahap *Pre Processing*, dalam tahap ini data yang tidak diperlukan akan dihilangkan seperti spasi, tanda baca, mengubah huruf kapital menjadi huruf kecil dan menghilangkan karakter-karakter yang tidak relevan, kemudian dilakukan tokenisasi dan melakukan stemming yaitu mengubah bentuk dari suatu kata menjadi kata dasar. Kemudian melakukan seleksi fitur dengan *term frequency* dan *inverse document frequency* untuk mengetahui suatu pola atau suatu bentuk term yang terdapat dalam perhitungan untuk membantu proses klasifikasi dalam suatu dokumen. Langkah selanjutnya yaitu menganalisa sentimen review film menggunakan perhitungan dengan metode *naïve bayes classifier* pada suatu review positif dan negatif dengan menambahkan *Laplace Smoothing*. tahap terakhir yaitu tahap pengujian dengan melakukan terhadap data testing menggunakan beberapa metode pengukuran yaitu *accuracy*, *recall*, *precision* dan *error rate.*

*C. **Metode Pengembangan Sistem***

a. Analisis Sistem
   Tahap analisis sistem merumuskan fungsionalitas dari sistem yang akan dibangun untuk digunakan dalam analisis sentiment review film. Juga analisa kebutuhan perangkat keras dan perangkat lunak yang akan digunakan.

e-ISSN: 2579-5570
p-ISSN: 2579-4116

b. Desain Sistem
Tahap desain membuat bagian ran sistem sesuai hasil analisa sistem. Mulai dari merancang basis data, merancang antarmuka atau *user interface* sampai dengan implementasi dan pengujian.

c. Implementasi
Implementasi aplikasi yang sudah siap dilakukan pada tahap ini dengan kriteria aplikasi yang digunakan dengan mudah dipahami oleh pengguna. Sistem aplikasi yang dibuat menggunakan *Framework Next.js* untuk sisi *frontend* dan *backend web*. Lalu untuk perhitungan metode *Naïve Bayes Classifier*, *text preprocessing dan tf-idf* menggunakan *python* dengan database nya menggunakan *MongoDB*.

IV. HASIL DAN PEMBAHASAN

Sistem yang dibangun merupakan sistem yang memanfaatkan metode *machine learning* untuk melakukan klasifikasi sentimen pada review film di sebuah situs perfilman yaitu IMDb. Hasil yang dikeluarkan berupa sentimen negatif atau positif mengenai suatu film. Metode *machine learning* yang digunakan yaitu *Naive Bayes* yang mempunyai tingkat akurasi yang besar berdasarkan penelitian-penelitan yang sudah dilakukan sebelumnya.

A. Analisis Sistem

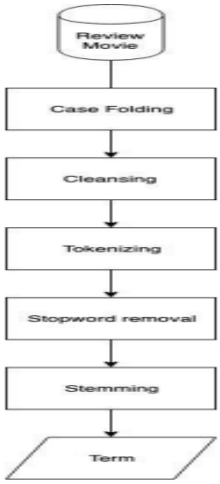
Sistem terdiri dari user admin. Dimana tugasnya adalah melakukan penginputan dataset berupa file excel yang sesuai dengan sample yang akan digunakan sebagai data training ke dalam sistem, berdasarkan label yang telah diketahui. Kemudian akan dilakukan proses *preprocessing*. Proses *preprocessing* dilakukan untuk membentuk model fitur training dari koleksi dokumen yang sudah diinputkan.

B. Preprocessing

Tahap awal yang dilakukan yaitu *preprocessing* data teks. Data yang digunakan berupa data kotor yang sudah diklasifikasikan menjadi dua kelas, yaitu kelas positif dan negatif. Proses *preprocessing* terdiri dari *case folding*, *cleansing*, *tokenizing*, *stopword removal*, dan *stemming*. Proses tersebut dapat dilihat pada Gambar 3.

a. *Case Folding*
Tahap *case folding* merupakan proses preprocessing awal yang mengubah huruf kapital menjadi huruf kecil pada semua teks yang ada dalam dataset latih.

b. *Cleansing*
Tahap *preprocessing* kedua yaitu *cleansing*, tahap membersihkan kata atau karakter pada teks yang tidak mempunyai pengaruh terhadap hasil klasifikasi sentimen. Kata atau karakter yang dibersihkan diantaranya simbol (@, #, %, &, !), emotikon, dan link URL.



Gambar 3 ***Preprocessing***

c. *Tokenizing*
Tahap *preprocessing* ketiga yaitu *tokenizing*, merupakan tahap untuk memisahkan teks menjadi kata. Tahap ini bertujuan agar mempermudah dalam melakukan pembobotan tiap kata.

d. *Stopword Removal*
Tahap *preprocessing* keempat yaitu *stopword removal*, tahap menghilangkan kata yang tidak sesuai dengan topik teks, bila tidak ada ataupun ada kata tersebut tidak mempengaruhi klasifikasi sentimen pada teks.

e. *Stemming*
Tahap akhir dari *preprocessing* yaitu proses *stemming*, merupakan proses untuk mengubah semua kata yang mempunyai imbuhan menjadi kata dasar.

*C. **Tahap Pelatihan Naïve Bayes***

Proses pelatihan digunakan untuk mendapatkan fitur latih yang dihasilkan dengan menggunakan metode *Naive Bayes* dan *Laplace Smoothing*. Tahap pelatihan terdiri dari preprocessing teks yang menghasilkan dataset, seleksi fitur menghasilkan fitur TF IDF, berikutnya proses penghitungan probabilitas *term* dan *laplace smoothing* yang menghasilkan fitur latih.

*D. **Tahap Pengujian Naïve Bayes***

Tahap pengujian dilakukan pada keseluruhan proses yang dilakukan pada tahap pelatihan. Proses pengujian bertujuan untuk menganalisis sentimen pada teks yang dimasukkan oleh user, dan untuk mengetahui nilai akurasi dari metode *Naive Bayes* dengan seleksi fitur menggunakan *Term frequency* (TF) dan *Inverse Document Frequency* (IDF). Nilai akurasi tersebut menunjukkan performa dari metode *Naive Bayes* dalam menganalisis sentimen terhadap review film.

e-ISSN: 2579-5570
p-ISSN: 2579-4116

E. Desain Database

Berikut adalah tabel-tabel yang terdapat dalam basis data sistem aplikasi analisis sentimen review film dengan algoritma *Naïve Bayes*.

a. Tabel Users
Tabel Users memuat data pengguna untuk dilakukan autentikasi pengguna saat proses masuk ke dalam sistem aplikasi. Untuk nama field dan tipe data ditunjukan pada Tabel 5.

Tabel 5 Tabel Users

Nama Field	Tipe Data	Keterangan
_id	ObjectId	Id user
Name	String	Nama pengguna
Photo	String	Foto profil pengguna
Username	String	Identifikasi nama pengguna
Password	String	Kata sandi pengguna
createdAt	Date	Tanggal saat data dibuat
updatedAt	Date	Tanggal saat data diubah

b. Tabel *Settings*
Tabel Settings memuat data untuk rasio perbandingan data pada *datasets* yang ditunjukan pada Tabel 6.

Tabel 6 Tabel Settings

Nama Field	Tipe Data	Keterangan
_id	ObjectId	Id user
firstRatio	Integer	Nilai rasio pertama
secondRatio	Integer	Nilai rasio kedua

c. Tabel *Datasets*
Tabel *Datasets* memuat data untuk dataset yang sudah berhasil di import seperti yang ditunjukan pada Tabel 7.

Tabel 7 Tabel Datasets

Nama Field	Tipe Data	Keterangan
_id	ObjectId	Id user
Review	String	Konten teks
Sentiment	String	Label / kelas teks

d. Tabel *Text Processing*
Tabel *Text Processing* memuat data hasil dari text preprocessing yang sudah diolah dari dataset seperti yang ditunjukan pada Tabel 8.

Tabel 8 Tabel *Text Processing*

Nama Field	Tipe Data	Keterangan
_id	ObjectId	Id user
Review	String	Konten teks
Sentiment	String	Label / kelas teks
TextProcessed	String	Hasil text preprocessing

e. Tabel *Terms*
Tabel *Terms* memuat data hasil dari perhitungan tf-idf yang sudah diolah dari data tabel *text processings*. Seperti yang ditunjukan pada Tabel 9.

Tabel 9 Tabel Terms

Nama Field	Tipe Data	Keterangan
_id	ObjectId	Id user
Word	String	Kata dari sebuah teks
Df	Integer	Term frequency
Idf	Double	Inverse document frequency

f. Tabel *Classifications*
Tabel *Classifications* memuat data uji hasil dari rasio perbandingan data yang sudah ditentukan. Seperti yang ditunjukan pada Tabel 10.

Tabel 10 Tabel Classifications

Nama Field	Tipe Data	Keterangan
_id	ObjectId	Id user
Review	String	Konten teks
Sentiment	String	Label / kelas teks
TextProcessed	String	Hasil text preprocessing

g. Tabel *Examinations*
Tabel *Examinations* memuat data hasil pengujian *confusion matrix* dari data uji dan data latih. Seperti yang ditunjukan pada Tabel 11.

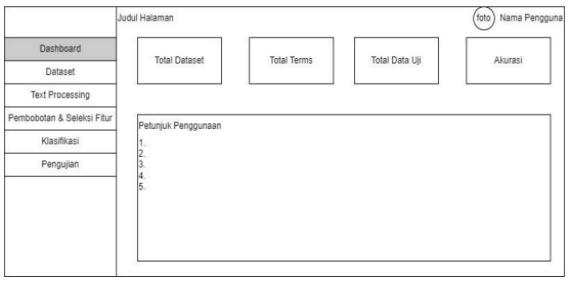
Tabel 11 Tabel Examinations

Nama Field	Tipe Data	Keterangan
_id	ObjectId	Id user
Accuracy	String	Persentase data uji
True_positive	Integer	Prediksi positif dan benar
False_negatif	Integer	Prediksi negatif dan salah
False_positive	Integer	Prediksi positif dan salah
True_negatif	Integer	Prediksi negatif dan benar

F. Desain Antarmuka

Desain antarmuka adalah tampilan halaman yang digunakan oleh sebuah aplikasi berinteraksi dengan pengguna agar mudah untuk memasukkan data. Pada penelitian ini penulis membuat halaman login, halaman dashboard, halaman dataset, halaman *text processing*, halaman pembobotan & seleksi fitur, halaman klasifikasi dan halaman pengujian.

a. Halaman dashboard
Rancangan halaman dashboard memuat tampilan antarmuka untuk memberikan pandangan sekilas tentang data yang ada dan petunjuk penggunaan sistem aplikasi. Dan digunakan sebagai indikasi bahwa pengguna sudah bisa masuk ke dalam sistem. Seperti yang ditunjukan pada Gambar 4 dibawah ini.

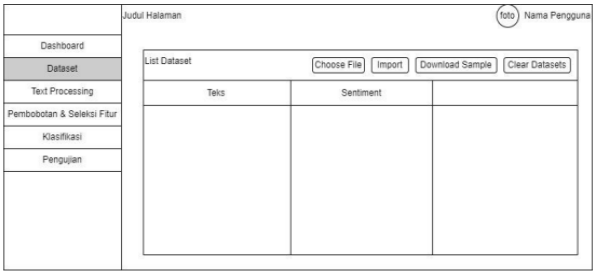


Gambar 4 Halaman dashboard

e-ISSN: 2579-5570
p-ISSN: 2579-4116

b. Halaman dataset

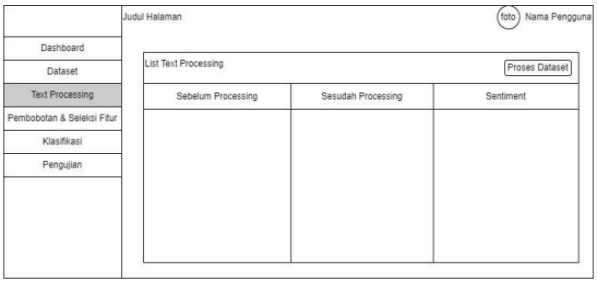
Rancangan halaman Dataset memuat tampilan daftar data berbentuk tabel, beberapa tombol yaitu Import, *Download Sample, Clear Datasets* dan input file untuk memasukkan data berbentuk csv agar bisa diolah oleh sistem untuk disimpan. Tombol Import berfungsi untuk memproses file yang sudah dimasukkan sebelumnya. Tombol Download Sample berfungsi untuk mengunduh contoh format dataset yang boleh di import. Sedangkan *Tombol Clear Datasets* berfungsi untuk menghapus keseluruhan data untuk dilakukan perhitungan ulang. Seperti yang ditunjukan pada Gambar 5 dibawah ini.



Gambar 5 Halaman dataset

c. Halaman *text processing*

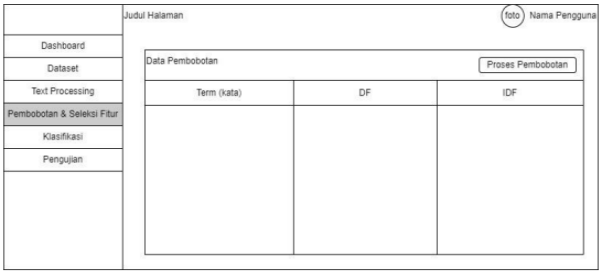
Rancangan halaman *Text Processing* memuat tampilan list data berbentuk tabel dan tombol Proses Dataset. Tombol Proses Dataset berfungsi untuk mengolah dataset yang sudah dimasukkan sebelumnya untuk dilakukan *text preprocessing*. Seperti yang ditunjukan pada Gambar 6 dibawah ini.



Gambar 6 Halaman *text processing*

d. Halaman pembobotan & seleksi fitur

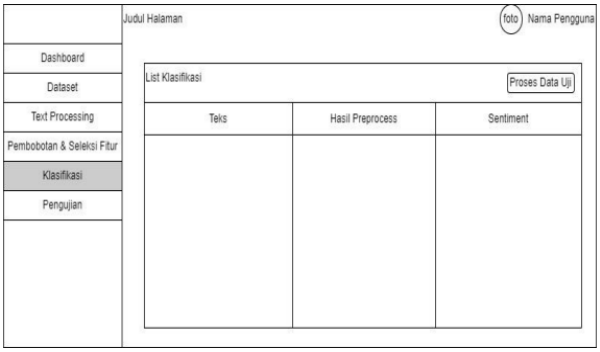
Rancangan halaman Pembobotan & Seleksi Fitur memuat tampilan list data berbentuk tabel dan tombol Proses Pembobotan. Tombol Proses Pembobotan berfungsi untuk mengolah dataset hasil dari text *preprocessing* yang sebelumnya sudah dilakukan untuk dilakukan perhitungan tf-idf. Seperti yang ditunjukan pada Gambar 7 dibawah ini.



Gambar 7 Halaman pembobotan & seleksi fitur

e. Halaman klasifikasi

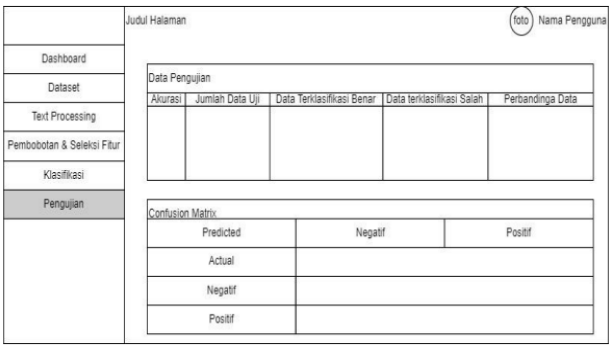
Rancangan halaman Klasifikasi memuat tampilan list data berbentuk tabel yang berisi data uji hasil dari rasio perbandingan data yang sudah ditentukan sebelumnya. Dan ada tombol Proses Data Uji untuk memproses perhitungan data uji dan data latih. Seperti yang ditunjukan pada Gambar 8 dibawah ini.



Gambar 8 Halaman klasifikasi

f. Halaman pengujian

Rancangan halaman Pengujian memuat tampilan list data berbentuk tabel yang berisi data pengujian akurasi dan *confusion matrix* hasil dari proses data uji pada halaman klasifikasi. Seperti yang ditunjukan pada Gambar 9 dibawah ini.

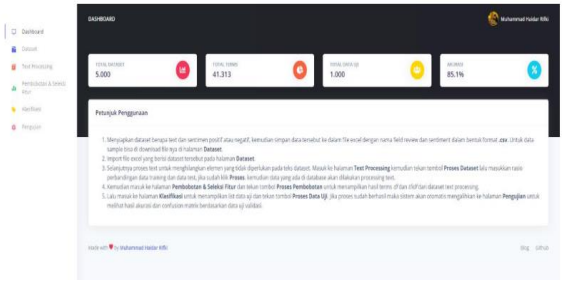


Gambar 9 Halaman pengujian

G. *Implementasi*

a. Halaman dashboard

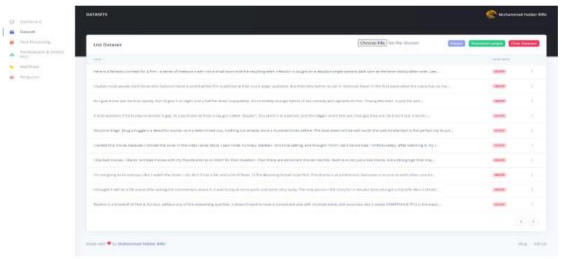
Halaman dashboard berisi data total dataset, total terms pada dataset, total data uji, akurasi dan petunjuk penggunaan aplikasi. Yang ditunjukan pada Gambar 10.



Gambar 10 Halaman dashboard

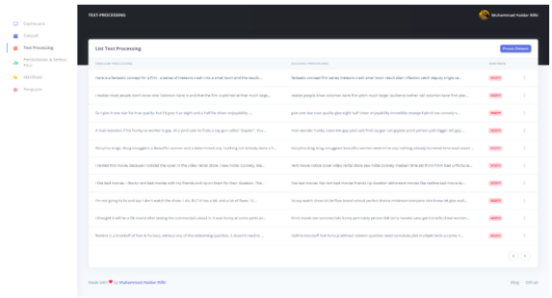
e-ISSN: 2579-5570
p-ISSN: 2579-4116

b. Halaman dataset
Halaman Dataset berisikan data review film yang telah diberi label. Admin bisa mengimport dataset dengan format file .csv atau mengunduh contoh data sample untuk membuat dataset. Terdapat tombol clear datasets untuk menghapus keseluruhan dataset dan memulai perhitungan ulang. Yang ditunjukkan pada Gambar 11.



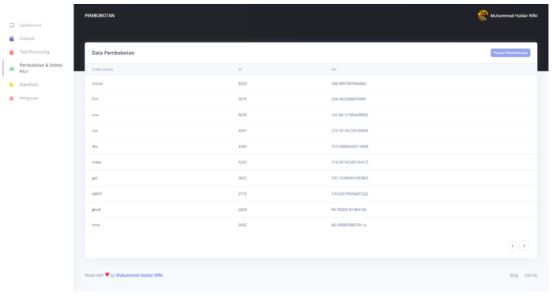
Gambar 11 Halaman dataset

c. Halaman *text processing*
Halaman *Text Processing* memuat hasil text pre-processing dataset yang telah dimasukkan. Pada halaman ini admin melakukan input rasio perbandingan data dan memproses text *pre-processing*. Tabel data pada halaman ini menampilkan data teks sebelum dan sesudah *pre-processing*, seperti ditunjukkan pada Gambar 12.



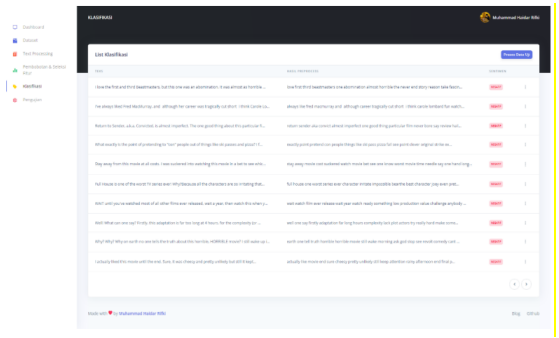
Gambar 12 Halaman text processing

d. Halaman pembobotan & seleksi fitur
Halaman Pembobotan & Seleksi Fitur memuat tentang data hasil setiap kata yang telah diberi pembobotan menggunakan metode tf-idf berdasarkan data hasil pre-processing sebelumnya. Yang ditunjukan pada Gambar 13.



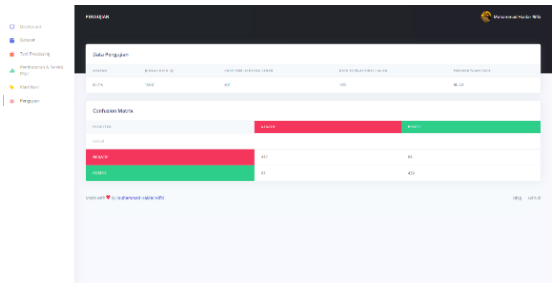
Gambar 13 Halaman pembobotan & seleksi fitur

e. Halaman klasifikasi
Halaman Klasifikasi memuat tabel data uji yang menampilkan data sesuai dengan rasio perbandingan data yang telah ditentukan. Yang ditunjukan pada Gambar 14.



Gambar 14 Halaman klasifikasi

f. Halaman Pengujian
Halaman pengujian memuat data hasil pengujian yang dilakukan pada penelitian ini yakni tabel hasil nilai akurasi pengujian dengan *confusion matrix*. Yang ditunjukan pada Gambar 15.



Gambar 15 Halaman pengujian

H. *Hasil Pengujian*

a. Black Box Testing
Black Box Testing untuk menguji fungsional sistem apakah sesuai dengan kebutuhan pengguna yang telah ditetapkan. Berikut hasil pengujian black box terdapat pada tabel 12.

Tabel 12 Black Box Testing

No	Keterangan	Pengujian	Hasil
1.	Halaman Login	Sistem hanya mengijinkan user yang terdaftar pada database untuk masuk ke sistem.	Sesuai
		Jika pengguna melakukan kesalahan penulisan username atau password maka tidak bisa masuk ke dalam sistem dan menampilkan pesan error	Sesuai
2.	Halaman Dataset	Dapat melakukan download sample dataset	Sesuai

e-ISSN: 2579-5570
p-ISSN: 2579-4116

No	Keterangan	Pengujian	Hasil
		Dapat melakukan import dataset yang sudah dibuat	Sesuai
		File yang bisa di import hanya yang berbentuk excel	Sesuai
		Dapat menampilkan dataset yang telah di import	Sesuai
3.	Halaman Text Processing	Dapat melakukan input rasio perbandingan data	Sesuai
		Dapat melakukan text pre-processing seperti stopword removal, cleansing, case folding, tokenizing, stemming.	Sesuai
		Dapat menampilkan data sebelum dan sesudah pre-processing.	Sesuai
4.	Halaman Pembobotan & Seleksi Fitur	Dapat melakukan proses pembobotan terms tf-idf.	Sesuai
		Dapat menampilkan data terms tf-idf hasil pembobotan.	Sesuai
5.	Halaman Pembobotan & Seleksi Fitur	Dapat melakukan proses pembobotan terms tf-idf.	Sesuai
		Dapat menampilkan data terms tf-idf hasil pembobotan.	Sesuai
6.	Halaman Pengujian	Dapat menampilkan data pengujian hasil akurasi dan confusion matrix.	Sesuai

b. Confusion Matrix

Pengujian dengan *Confusion Matrix* untuk model Algoritma Multinomial *Naïve Bayes* adalah sebagai berikut yang ditunjukan pada Gambar 17.

Gambar 17 Confusion Matrix

Kemudian dilakukan uji performa untuk algoritma *Naïve Bayes Classifier* seperti berikut.

$$Accuracy = \frac{(98+113)}{(98+23+8+113)} = 0{,}871 \times 100 = 87{,}1\%$$

$$Recall = \frac{(98)}{(98+8)} = 0{,}924 \times 100 = 92{,}4\%$$

$$Precision = \frac{(98)}{(98+23)} = 0{,}809 \times 100 = 80{,}9\%$$

$$Error\ Rate = \frac{(23+8)}{(98+23+8+113)} = 0{,}128 = 12{,}8\%$$

Dari perhitungan uji performa hasil implementasi algoritma *Naïve Bayes* diatas maka didapatkan hasilnya yaitu *Accuracy* 87,2%, *Recall* 92,4%, *Precision* 80,9% dan *Error Rate* 12,8%.

V. KESIMPULAN

A. *Kesimpulan*

Dari sistem sentimen analisis review film menggunakan *metode naive bayes classifier* yang sudah dibuat dengan jumlah dataset 1216 yang terbagi menjadi dua kelas sentimen berupa 899 sentimen positif dan 317 sentimen negatif dan dari total 1216 dataset dengan rasio perbandingan data 80:20 menjadi sebanyak 974 data latih dan 242 data uji. Maka diperoleh hasil akurasi 87.2%. Dari 242 data uji sebanyak 98 diprediksikan sesuai yaitu "Negatif" dan 23 data diprediksikan "Positif" ternyata "Negatif". dan sebanyak 8 data sebanyak diprediksi "Negatif" ternyata "Positif" dan sebanyak 113 data diprediksikan sesuai yaitu "Positif".

B. ***Saran***

Menurut penulis ketika melakukan penelitian klasifikasi sentimen review film pada forum IMDb dengan menggunakan metode *Naïve Bayes Classifier*, penulis menemukan beberapa kendala dan kekurangan selama proses pembuatan sistem.

Adapun saran untuk penelitian selanjutnya adalah sebagai berikut :

a. Pada penelitian selanjutnya dapat dikembangkan dengan menggunakan sumber data dari platform lain seperti Facebook, Instagram, Twitter, Youtube, dll.
b. Dalam penelitian ini, metode klasifikasi yang dipakai adalah *Naïve Bayes Classifier*. Penelitian selanjutnya dapat menggunakan metode klasifikasi lain seperti metode *K-Nearest Neighbor, Modified K-Nearest Neighbor, C45, random forest, dan Support Vector Machine (SVM).* Hal ini dapat dilakukan untuk melihat perbandingan variasi hasil akurasi berbagai metode klasifikasi.

DAFTAR PUSTAKA

[1] H. Tuhuteru, “Analisis Sentimen Masyarakat terhadap Pembatasan Sosial Berksala Besar Menggunakan Algoritma Support Vector Machine,” *Inf. Syst. Dev.*, vol. 5, no. 2, pp. 7–13, 2020.

[2] C. Huda and M. Betty Yel, “Analisa Sentimen Tentang Ibu Kota Nusantara (IKN) Dengan Menggunakan Algoritma K-Nearest Neighbors (KNN) dan Naïve Bayes,” *J. Ilmu Komput. dan Sist. Inf.*, vol. 7, no. 1, pp. 126–130, 2024, doi: 10.55338/jikomsi.v7i1.2846.

[3] I. Iwandini, A. Triayudi, and G. Soepriyono, “Analisa Sentimen Pengguna Transportasi Jakarta Terhadap Transjakarta Menggunakan Metode Naives Bayes dan K-Nearest Neighbor,” *J. Inf. Syst. Res.*, vol. 4, no. 2, pp. 543–550, 2023, doi: 10.47065/josh.v4i2.2937.

[4] A. D. Pratama and H. Hendry, “Analisa Sentimen Masyarakat Terhadap Penggunaan Chatgpt Menggunakan Metode Support Vector Machine (Svm),” *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 9, no. 1, pp. 327–338, 2024, doi:

e-ISSN: 2579-5570
p-ISSN: 2579-4116

10.29100/jipi.v9i1.4285.

[5] I. S. K. Idris, Y. A. Mustofa, and I. A. Salihi, "Analisis Sentimen Terhadap Penggunaan Aplikasi Shopee Mengunakan Algoritma Support Vector Machine (SVM)," *Jambura J. Electr. Electron. Eng.*, vol. 5, no. 1, pp. 32–35, 2023, doi: 10.37905/jjeee.v5i1.16830.

[6] E. Suryati, Styawati, and A. A. Aldino, "Analisis Sentimen Transportasi Online Menggunakan Ekstraksi Fitur Model Word2vec Text Embedding Dan Algoritma Support Vector Machine (SVM)," *J. Teknol. Dan Sist. Inf.*, vol. 4, no. 1, pp. 96–106, 2023.

[7] I. Adiwijaya, "Text Mining dan Knowledge Discovery," *Kolok. bersama komunitas datamining Indones. soft-computing Indones.*, pp. 1–9, 2006.

[8] Sugiyono, *Sugiyono, Metode Penelitian dan Pengembangan Pendekatan Kualitatif, Kuantitatif, dan R&D , (Bandung: Alfabeta, 2015), 407 1*. Bandung: ALFABETA, cv., 2015.

[9] M. Ridwan, H. Suyono, and M. Sarosa, "Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," *Eeccis*, vol. 7, no. 1, pp. 59–64, 2013, doi: 10.1038/hdy.2009.180.

[10] F. Nurhuda, S. Widya Sihwi, and A. Doewes, "Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier," *J. Teknol. Inf. ITSmart*, 2016, doi: 10.20961/its.v2i2.630.

[11] A. Jananto, "Algoritma Naive Bayes untuk Mencari Perkiraan Waktu Studi Mahasiswa," *Teknol. Inf. Din.*, vol. 18, no. 1, pp. 9–16, 2013.

[12] M. S. Mustafa, M. R. Ramadhan, and A. P. Thenata, "Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," *Creat. Inf. Technol. J.*, vol. 4, no. 2, p. 151, 2018, doi: 10.24076/citec.2017v4i2.106.

[13] A. Radili and S. Sanjaya, "Penerapan Metode Winnowing Fingerprint dan Naive Bayes untuk Pengelompokan Dokumen," *J. CoreIT J. Has. Penelit. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, p. 69, 2018, doi: 10.24014/coreit.v3i2.4418.

[14] T. Informatika, U. Malikussaleh, and A. Utara, "Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi," *J. Inform. Ahmad Dahlan*, vol. 8, no. 1, p. 102632, 2014, doi: 10.26555/jifo.v8i1.a2086.

[15] R. R. Setiawan *et al.*, "Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, 2017, doi: 10.1074/jbc.M209498200.

[16] T. Informatika, U. Malikussaleh, and A. Utara, "Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi," vol. 8, no. 1, pp. 884–898, 2014, doi: 10.26555/jifo.v8i1.a2086.

[17] I. Artikel, "Pemanfaatan Naïve Bayes Untuk Merespon Emosi Dari Kalimat Berbahasa Indonesia," *Unnes J. Math.*, vol. 4, no. 2, 2015, doi: 10.15294/ujm.v4i2.9706.