TESIS

KLASIFIKASI GENRE ANIME BERDASARKAN SINOPSIS MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBORS DAN NAIVE BAYES

Disusun oleh:

Nama	**: Gede Putra Aditya Brahmantha**
NIM	**: 22.55.1195**
Konsentrasi	**: Business Intelligence**

PROGRAM STUDI S2 INFORMATIKA

PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2024

TESIS

KLASIFIKASI GENRE ANIME BERDASARKAN SINOPSIS MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBORS DAN NAIVE BAYES

ANIME GENRE CLASSIFICATION BASED ON SYNOPSIS USING K-NEAREST NEIGHBORS AND NAIVE BAYES ALGORITHMS

Diajukan untuk memenuhi salah satu syarat memperoleh derajat Magister

Disusun oleh:

Nama	**: Gede Putra Aditya Brahmantha**
NIM	**: 22.55.1195**
Konsentrasi	**: Business Intelligence**

PROGRAM STUDI S2 INFORMATIKA

PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2024

HALAMAN PENGESAHAN

KLASIFIKASI GENRE ANIME BERDASARKAN SINOPSIS MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBORS DAN NAIVE BAYES

ANIME GENRE CLASSIFICATION BASED ON SYNOPSIS USING K-NEAREST NEIGHBORS AND NAIVE BAYES ALGORITHMS

Dipersiapkan dan Disusun oleh

Gede Putra Aditya Brahmantha

22.55.1195

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Kamis, 11 Juli 2024

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 11 Juli 2024
Rektor

Prof. Dr. M. Suyanto, M.M.
NIK. 190302001

HALAMAN PERSETUJUAN

KLASIFIKASI GENRE ANIME BERDASARKAN SINOPSIS MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBORS DAN NAIVE BAYES

ANIME GENRE CLASSIFICATION BASED ON SYNOPSIS USING K-NEAREST NEIGHBORS AND NAIVE BAYES ALGORITHMS

Dipersiapkan dan Disusun oleh

Gede Putra Aditya Brahmantha

22.55.1195

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Kamis, 11 Juli 2024

Pembimbing Utama

Prof. Dr. Ema Utami, S.Si., M.Kom
NIK. 190302037

Anggota Tim Penguji

Tonny Hidayat, S.Kom., M.Kom., Ph.D.
NIK. 190302182

Pembimbing Pendamping

Ainul Yaqin, M.Kom
NIK. 190302255

Dhani Ariatmanto, S.Kom., M.Kom., Ph.D.
NIK. 190302197

Prof. Dr. Ema Utami, S.Si., M.Kom
NIK. 190302037

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 11 Juli 2024
Direktur Program Pascasarjana

Prof. Dr. Kusrini, M.Kom
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Gede Putra Aditya Brahmantha
NIM : 22.55.1195
Konsentrasi : Business Intelligence

Menyatakan bahwa Tesis dengan judul berikut:
KLASIFIKASI GENRE ANIME BERDASARKAN SINOPSIS MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBORS DAN NAIVE BAYES

Dosen Pembimbing Utama : Prof. Dr. Ema Utami, S.Si., M.Kom
Dosen Pembimbing Pendamping : Ainul Yaqin, M.Kom

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 11 Juli 2024
Yang Menyatakan,

Gede Putra Aditya Brahmantha

HALAMAN PERSEMBAHAN

Puji Syukur penulis panjatkan kepada Tuhan Yang Maha Esa, yang telah memberikan kesehatan, rahmat dan hidayah, sehingga penulis diberikan kesempatan untuk menyelesaikan tesis ini, sebagai salah satu syarat untuk mendapatkan gelar magister. Walaupun jauh dari kata sempurna, namun penulis bangga telah mencapai pada titik ini, yang akhirnya tesis ini bisa selesai diwaktu yang tepat.

Tesis ini saya persembahkan untuk :

1. Kedua Orang Tua Tercinta, Ibu Ratna dan Bapak Sudarsana yang terus mendoakan untuk kelancaran studi.
2. Adik tersayang, Indranatha yang telah memberikan semangat dan dukungan moral.
3. Xavero, Stefanie, dan Alicia sebagai Teman terdekat yang setia menemani dalam segala situasi dan selalu mendukung Saya.
4. TIM ITE RSUD Praya, Bapak Jihad, Bapak Ilham, Bapak Haerul, Bapak Huda, dan Ibu Ajeng yang memberikan semangat dan motivasi.

HALAMAN MOTTO

"Per Aspera Ad Astra"

"Menuju bintang melalui jerih payah"

KATA PENGANTAR

Penelitian dengan judul **Klasifikasi Genre Anime Berdasarkan Sinopsis Menggunakan Algoritma K-Nearest Neighbors Dan Naive Bayes** ini disusun dalam pelaksanaan Tesis pada Program Studi Magister Informatika Universitas Amikom. Tesis ini disusun dengan harapan dapat menjadi pedoman dan arahan dalam melaksanakan penelitian di atas.

Sehubungan dengan telah terselesaikannya penelitian ini, maka diucapkan terimakasih dan penghargaan kepada berbagai pihak yang telah membantu penyusun, antara lain :

1. Ibu Prof. Dr. Ema Utami, S.Si., M.Kom sebagai Pembimbing Utama yang telah banyak membantu menyempurnakan penelitian ini;
2. Bapak Ainul Yaqin, M.Kom sebagai Pembimbing Pendamping yang telah banyak membantu menyempurnakan penelitian ini;

Penulis mengharapkan tesis ini dapat memberikan sumbangsih bagi pendidikan yang selalu berkembang seiring dengan tuntutan zaman.

Yogyakarta, 11 Juli 2024

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR ISTILAH

FN : False Negative

FP : False Positive

HTML : Hypertext Markup Language

KNN : K-Nearest Neighbors

LR : Logistic Regression

MAL : MyAnimeList

MI : Mutual Information

NB : Naïve Bayes

NBC : Naïve Bayes Classifier

NLP : Natural Language Processing

SVM : Support Vector Machine

TF-IDF : Term Frequency Inverse Document Frequency

TN : True Negative

TP : True Positive

URL : Uniform Resource Locator

INTISARI

Anime adalah salah satu bentuk hiburan populer yang berupa animasi yang berasal dari Jepang, dengan popularitas anime, banyak streaming services yang menyediakan anime dalam konten layanan mereka. Anime memiliki sebuah cerita layaknya film dan memiliki banyak genre, Genre adalah istilah yang dipakai untuk mengelompokkan media ke dalam kategori-kategori yang memiliki ciri-ciri yang serupa. Saat ini, pengelompokkan genre anime masih dilakukan secara manual oleh penerbit dan perlu membaca seluruh sinopsis atau menonton anime tersebut, mengakibatkan investasi waktu yang signifikan. Untuk menyelesaikan masalah tersebut peneliti mengusulkan klasifikasi genre anime berdasarkan sinopsis menggunakan algoritma K-Nearest Neighbors dan Naive Bayes dan membandingkan kinerja algoritma tersebut beserta penggunaan kombinasi fitur yang berbeda.

Dalam penelitian ini dilakukan klasifikasi genre anime berdasarkan sinopsis berbahasa Inggris. Klasifikasi genre dilakukan untuk mengelompokkan genre menjadi 4 jenis yaitu fantasy, mystery, romance dan sports. Setiap genre berisi 250 data sinopsis sehingga total dataset mencapai 1000 sinopsis. Dilakukan tahap preprocessing dan pembobotan TF-IDF serta seleksi fitur Mutual Information yang dilanjutkan ke tahap klasifikasi K-Nearest Neighbors dan Naive Bayes.

Hasil terbaik didapatkan menggunakan algoritma Naive Bayes dengan menggunakan seluruh kombinasi fitur preprocessing tanpa melibatkan seleksi fitur mutual information dengan hasil nilai akurasi 78.0%, precision 78.087%, recall 78.0%, F1-score 77.902% menungguli hasil terbaik yang dihasilkan algoritma K-Nearest Neighbors dengan menggunakan seluruh kombinasi fitur preprocessing tanpa melibatkan seleksi fitur mutual information dengan hasil akurasi 78.0%, precision 78.087%, recall 78.0%, F1-score 77.902%. Hal ini membuktikan bahwa algoritma Naive Bayes memiliki kinerja yang lebih baik dari K-Nearest Neighbors dalam melakukan Klasifikasi genre anime berdasarkan sinopsis.

Kata kunci: anime, genre, machine learning, k-nearest neighbors, klasifikasi

ABSTRACT

Anime is a popular form of entertainment in the form of animation originating from Japan. With the popularity of anime, many streaming services include anime in their contents. Anime has a story like a movie and comes in various genres. Genre is a term used to categorize media into groups that share similar characteristics. Currently, anime genre classification is still done manually by publishers, requiring them to read the entire synopsis or watch the anime, resulting in significant time investment. To address this issue, the researcher proposed classifying anime genres based on synopsis using K-Nearest Neighbors and Naive Bayes algorithms and compare their performance along with the use of different feature combinations.

In this research, anime genre classification performed based on English-language synopses. Genre classification is done to classify genres into 4 types, namely fantasy, mystery, romance and sports. Each genre contains 250 synopses data bringing the total dataset to 1,000 synopses. Preprocessing and TF-IDF weighting and Mutual Information feature selection were carried out, followed by K-Nearest Neighbors and Naive Bayes classification.

The best results were obtained using Naive Bayes algorithm with all preprocessing feature combinations without involving mutual information feature selection, achieving an accuracy 78.0%, precision 78.087%, recall 78.0%, and F1-score 77.902% outperforming the best results from K-Nearest Neighbors algorithm using all preprocessing feature combinations without mutual information feature selection with an accuracy 78.0%, precision 78.087%, recall 78.0%, and F1-score 77.902%. This proves that Naive Bayes algorithm outperforms K-Nearest Neighbors in classifying anime genres based on synopsis.

Keyword: anime, genre, machine learning, k-nearest neighbors, classification

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Industri hiburan adalah salah satu bidang yang sangat banyak dinikmati oleh banyak orang. Industri ini meliputi film, acara TV, musik, game, dan lain-lain. Anime, gaya animasi Jepang, adalah salah satu fasilitas hiburan terbesar yang dinikmati oleh banyak orang. Dengan berkembangnya Teknologi & Informasi, anime dengan sangat mudahnya diakses oleh banyak orang melalui Streaming Services seperti Netflix, BiliBili, dan Crunchyroll. Anime memiliki sebuah cerita layaknya film, anime juga memiliki banyak genre cerita. Biasanya sebelum memilih anime untuk diputar, seseorang akan melihat sebuah sinopsisnya terlebih dahulu. Sinopsis merupakan ringkasan atau garis besar naskah yang menggambarkan isi dari sebuah film, buku, atau pementasan yang dilakukan baik secara konkrit maupun secara abstrak. Genre merupakan istilah yang digunakan untuk mengklasifikasikan teks media ke dalam kelompok-kelompok dengan karakteristik sejenis (Rayner et al., 2004). Saat ini, Seseorang harus membaca sinopsis atau menonton suatu Anime untuk mengetahui genre pada Anime tersebut, begitu pula dengan seseorang yang mengklasifikasikan secara manual untuk dimuat ke dalam situs seperti Streaming Services. Karena itu, mengklasifikasikan genre atau mengelompokkan anime adalah metode untuk mengidentifikasi keterkaitan antara setiap anime, sehingga memudahkan penonton menemukan anime yang sesuai dengan preferensi mereka.

Penelitian sebelumnya yang terkait klasifikasi genre dilakukan oleh (Saputra et al., 2019), Dalam penelitian tersebut data teks sinopsis pada film Indonesia dapat digunakan sebagai fitur untuk menentukan genre film yang sesuai dengan memanfaatkan algoritma machine learning. Model yang digunakan menunjukkan bahwa hasil terbaik adalah SVM+TF-IDF Classifier menggunakan unigram dengan skor f1 sebesar 45%. Terdapat penelitian lain yang membahas klasifikasi berdasarkan genre film yang dilakukan oleh (Akbar et al., 2023), penelitian tersebut bersifat klasifikasi multi-label dan digunakan Algoritma Support Vector Machine, Regresi Logistik, dan Naive Bayes dengan hasil pengujian terbaik yaitu SVM. Selain itu pula terdapat penelitian yang melakukan pengujian Klasifikasi Sinopsis Novel Menggunakan Metode Naïve Bayes (Rahmayanti et al., 2019) yang menghasilkan akurasi sebesar 80.5%. Namun Penelitian terkait sebelumnya membahas genre film dan novel bukan anime.

Naive Bayes Classifier adalah algoritma klasifikasi yang menghitung probabilitas suatu peristiwa terjadi. Probabilitas dihitung secara independen dari fitur-fitur lain yang berkontribusi pada pengambilan keputusan. Dalam sederhana, pengklasifikasi Naive Bayes mengasumsikan bahwa kehadiran fitur tertentu dalam kelas tidak terkait dengan kehadiran fitur lainnya. Terlepas dari apakah ini fitur-fitur ini bergantung pada kehadiran satu sama lain, fitur-fitur ini berkontribusi pada probabilitas. Algoritma Naïve Bayes dipilih karena keunggulannya yang hanya membutuhkan sedikit pelatihan data untuk estimasi parameter yang diperlukan untuk klasifikasi. Algoritma Naive Bayes memberikan probabilitas bersyarat, yang bekerja berdasarkan teorema Bayes (Banlawe et al., 2021). Teorema Bayesian

digunakan untuk metode supervised learning serta metode statistik untuk memperkirakan klasifikasi dan model probabilitas yang mendasarinya dan memungkinkan untuk menangkap ketidakpastian tentang model secara teoritis dengan menentukan hasil probabilitas dengan demikian algoritma ini dapat memecahkan masalah diagnostik dan masalah prediksi (Patel & Parikh, 2020). K-Nearest Neighbors (KNN) adalah salah satu metode klasifikasi untuk sekumpulan data berdasarkan mayoritas kategori dan bertujuan untuk mengklasifikasikan objek baru berdasarkan atribut dan sampel dari data training. (Putra et al., 2022) Kelebihan dari metode K-Nearest Neighbors adalah dapat diaplikasikan pada data yang besar secara efektif dengan hasil yang akurat. (Permana et al., 2021) dan K-NN juga memiliki kelebihan yaitu pelatihan yang sangat cepat dan sederhana sehingga mudah dipelajari (Iriantoro et al., 2017).

Bedasarkan permasalahan yang ada dan penelitian – penelitian terkait yang menjadi dasar dalam melakukan penelitian, Maka peneliti menggunakan Naive Bayes dan K-Nearest Neighbors untuk mengetahui perbandingan hasil pengujian dari kedua algoritma tersebut dalam melakukan klasifikasi dan mengetahui kombinasi fitur yang menghasilkan pengujian terbaik dalam mengklasifikasikan genre anime berdasarkan sinopsis.

1.2. Rumusan Masalah

a. Berapa tingkat Recall, Precision, F-Measure dari Algoritma Naive bayes dan K-Nearest Neighbors dalam Mengklasifikasikan genre Anime berdasarkan sinopsis?

b. Apa kombinasi Fitur preprocessing terbaik untuk mengklasifikasikan Anime berdasarkan sinopsis menggunakan Algoritma Naive Bayes dan K-Nearest Neighbors?

c. Apakah menerapkan seleksi fitur Mutual Information dapat meningkatkan akurasi pada Algoritma Naive Bayes dan K-Nearest Neighbors?

1.3. Batasan Masalah

a. Hanya menggunakan dataset sinopsis anime yang berbahasa Inggris

b. Sinopsis yang digunakan hanyalah sinopsis dari anime itu sendiri, tidak termasuk adaptasi dari/ke novel, manga, film.

c. Program akan berjalan luar jaringan (offline)

d. Data yang digunakan dalam penelitian ini hanya berupa teks

e. Dataset diambil dari website MyAnimeList (MAL) dengan menggunakan teknik scraping

f. Melakukan Klasifikasi menggunakan algoritma Naive Bayes dan K-Nearest Neighbors

g. Melakukan perubahan kombinasi fitur preprocessing

h. Klasifikasi yang diterapkan adalah multi-class classification, yang artinya mengklasifikasikan data ke dalam salah satu dari tiga kelas atau lebih

i. Penelitian ini menggunakan Seleksi Fitur yaitu Mutual Information

1.4. Tujuan Penelitian

a. Evaluasi hasil klasifikasi yang telah dilakukan menggunakan confusion matrix

b. Menemukan Kombinasi Fitur terbaik untuk mengklasifikasikan Anime berdasarkan sinopsis menggunakan Algoritma Naive Bayes dan K-Nearest Neighbors
c. Mengetahui performa penerapan seleksi fitur yaitu Mutual Information dalam klasifikasi menggunakan Naive Bayes dan K-Nearest Neighbors

1.5. Manfaat Penelitian

a. Diharapkan dapat mengetahui akurasi dan efisiensi perubahan kombinasi Fitur terbaik untuk mengklasifikasikan Genre Anime berdasarkan sinopsis menggunakan Algoritma Naive Bayes dan K-Nearest Neighbors
b. Mengetahui penggunaan Seleksi fitur yaitu Mutual Information dalam memilih fitur terbaik dalam klasifikasi genre anime menggunakan Naive Bayes dan K-Nearest Neighbors
c. Bagi pengguna dapat digunakan untuk memberikan sistem rekomendasi anime berdasarkan sinopsis

BAB II

TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Penelitian tentang klasifikasi genre dan Algoritma Naive bayes telah banyak diteliti sebelumnya seperti Klasifikasi Multi Label genre film berdasarkan sinopsis (Akbar et al., 2023). Dalam penelitian tersebut, Algoritma klasifikasi multi-label yang digunakan adalah Algoritma Support Vector Machine, Regresi Logistik, dan Naive Bayes. Pencarian parameter optimal menggunakan GridSearch dari masing-masing algoritma. Hasil optimal pada penelitian ini diperoleh nilai f1-score sebesar 0.58 menggunakan algoritma SVM dengan ekstraksi fitur TF-IDF dengan dataset stemming, diikuti oleh NB dengan nilai f1-score sebesar 0.48 dan LR dengan nilai f1-score sebesar 0.43.

Ada pula penelitian lain yang membahas Klasifikasi Genre film berdasarkan sinopsis dari film indonesia (Saputra et al., 2019). Dalam penelitian tersebut, algoritma mampu menghasilkan klasifikasi teks yang baik dengan melihat nilai akurasi dan F1 score yang terbaik. Pada klasifikasi genre film Indonesia berdasarkan ringkasan sinopsis film Indonesia, algoritma klasifikasi Support Vector Machines dengan ekstraksi TF-IDF mampu menemukan nilai akurasi dan akurasi dan F1 score terbaik setelah data training (45%). Pada penelitian ini, masih terdapat kekurangan dan kesalahan pada hasil klasifikasi genre film saat melakukan input ringkasan sinopsis. Hal ini disebabkan karena kurangnya dataset terutama melihat jumlah film Indonesia dengan ringkasan sinopsis dan genre film yang masih

dapat digunakan dalam dataset. Perlunya penambahan dataset dan pemilihan data film Indonesia dengan kriteria bahasa Indonesia yang baik dan benar serta menyesuaikan genre film yang ada.

Terdapat Penelitian lainnya yang membahas tentang performa naive bayes dan KNN dalam klasifikasi sentimen terhadap penilaian jasa (Hermansyah & Sarno, 2020) klasifikasi K-NN memiliki akurasi yang lebih baik dibandingkan yang lain. Hasil yang diperoleh menunjukkan akurasi dari TextBlob, Naive Bayes dan K-NN sebesar 54.67%, 69.44%, dan 75%. Metode Textblob memiliki performa yang lebih buruk dibandingkan Naive Bayes dan K-NN dalam proses recall karena tipe pengklasifikasi yang tidak supervised type classifier dibandingkan dengan machine learning berbasis classifier berbasis machine learning sebesar 13,68%, 82,76%, dan 56,32%. Akan tetapi dalam proses recall Namun dalam proses recall memberikan keuntungan karena Textblob memaksa untuk menghitung seluruh dataset sebagai data uji memberikan hasil presisi sebesar TextBlob, Naive Bayes, dan K-NN sebesar 94.12%, 64.29%, dan 87.50%. Metode K-NN berkinerja lebih baik karena kemampuannya untuk menemukan kesamaan antara pengamatan dan sifat yang melekat untuk mengoptimalkan secara lokal. Peningkatan lebih lanjut dapat dilakukan dengan menggunakan dataset yang lebih besar dan lebih kompleks untuk meningkatkan nilai akurasi dalam analisis sentimen. Penelitian di masa depan diharapkan juga dapat memasukkan setiap aspek dari produk untuk meningkatkan hipotesis sentimen.

Penelitian selanjutnya membahas Klasifikasi Naive Bayes dalam melakukan klasifikasi laporan gangguan listrik (Fathoni et al., 2020), Dalam

penelitian tersebut, Berdasarkan penelitian yang telah dilakukan dengan menggunakan algoritma Naive Bayes Classifier dalam text mining diketahui bawah laporan yang teklasifikasi dibawah 180 menit ada 64,4% , 34,8% untuk laporan yang terklasifikasi kurang dari 1 hari dan 0,08% untuk laporan yang terklasifikasi lebih dari 1 hari. Perlunya penelitian terkait seberapa besar pengaruh akurasi terhadap Penggunaan stopwords dan stemming dalam preprocessing.

Selain itu terdapat penelitian Klasifikasi menggunakan Naive Bayes dalam mengklassifikasi komentar (Chatrina, Siregar et al., 2020). Untuk menentukan kategori dilakukan pengujian terhadap 50 data komentar dengan menggunakan algoritma Naïve Bayes Classifier didapatkan bahwa hasil nilai akurasi yaitu sebesar 68%. Dalam pengembangan selanjutnya, agar dapat melakukan penambahan data training (data latih) dalam jumlah yang lebih banyak dengan tujuan menghasilkan tingkat akurasi yang lebih tinggi karena semakin banyak data latih yang digunakan maka akan semakin akurat pula hasil klasifikasi dan dalam pengolahan komentar dapat menggunakan bahasa asing seperti bahasa inggris dan lainnya.

Penelitian lainnya menggunakan Algoritma Naive Bayes dalam melakukan klasifikasi keluhan masyarakat (Ariyanti & Iswardani, 2020). Dalam penelitian tersebut, Pengolahan data keluhan masyarakat ini melalui beberapa tahapan teks mining yaitu token, filter, stemming dan analyzing. Setelah melalui tahapan praproses, data tersebut akan dilakukan klasifikasi menggunakan algoritma Naïve Bayes, hasil perhitungan tersebut yang nantinya akan menunjukkan hasil kelas dari setiap data keluhan masyarakat yang masuk baik melalui telepon, sms. Penelitian

ini menghasikan tingkat akurasi mencapai 95%, sehingga dapat mengklasifikasikan keluhan masyarakat tiap-tiap instansi di pemerintah Kota Probolinggo.

Penelitian lainnya adalah mengkomparasi algoritma KNN, Naive Bayes, dan SVM dalam melakukan klasifikasi genre film berdasarkan sinopsis (Buslim et al., 2022). Pada penelitian tersebut, para peneliti mencoba membandingkan algoritma KNN, Naive Bayes, dan SVM untuk klasifikasi teks dengan masalah klasifikasi genre film berdasarkan sinopsis menggunakan dataset yang diperoleh dari Kaggle.com dan IMDB Dataset. Hasil dari penelitian ini menunjukkan bahwa dari 12 kali percobaan dengan masing-masing algoritma diuji sebanyak 4 kali, Support Vector Machine (SVM) merupakan algoritma yang memiliki performa paling baik dengan akurasi sebesar 90%, 93%, 65%, dan 63%.

2.2. Keaslian Penelitian

Tabel 2.1. Matriks literatur review dan posisi penelitian
KLASIFIKASI GENRE ANIME BERDASARKAN SINOPSIS MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBORS DAN NAIVE BAYES

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	Multi-Label Classification of Film Genres Based on Synopsis Using Support Vector Machine, Logistic Regression and Naïve Bayes Algorithms	J. Akbar, E. Utami and A. Yaqin, 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, Indonesia, 2022	Menguji dan Membandingkan klasifikasi multi-label dari Genre film berdasarkan sinopsis menggunakan algoritma SVM, LR, dan Naive Bayes	Algoritma klasifikasi multi-label yang digunakan adalah Algoritma Support Vector Machine, Regresi Logistik, dan Naive Bayes. Pencarian parameter optimal menggunakan GridSearch dari masing-masing algoritma. Hasil optimal pada penelitian ini diperoleh nilai f1-score sebesar 0.58 menggunakan algoritma SVM dengan ekstraksi fitur TF-IDF dengan dataset stemming, diikuti oleh	Peningkatan lebih lanjut lebih lanjut pada domain dan data set dengan menerapkan pendekatan ekstraksi fitur yang lebih spesifik untuk masalah klasifikasi genre film sehingga dapat sehingga dapat meningkatkan hasil f1-score.	Penelitian tersebut menggunakan synopsis film sebagai objek penelitian, penekanan penelitian ini adalah perbandingan algoritma klasifikasi yang diugunakan untuk mengklasifikasi genre secara multi-label.

Tabel 2.1. Lanjutan Matriks literatur review dan posisi penelitian

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
				NB dengan nilai f1-score sebesar 0.48 dan LR dengan nilai f1-score sebesar 0.43.		
2	The Classification of the Movie Genre based on Synopsis of the Indonesian Film	A. C. Saputra, A. B. Sitepu, Stanley, P. W. P. Yohanes Sigit, P. G. Sarto Aji Tetuko and G. C. Nugroho. 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), Yogyakarta, Indonesia, 2019, pp. 201-204	Menguji algoritma SVM dalam melakukan klasifikasi genre film indonesia berdasarkan sinopsis	Berdasarkan penelitian ini, algoritma mampu menghasilkan klasifikasi teks yang baik dengan melihat nilai akurasi dan F1 score yang terbaik. Pada klasifikasi genre film Indonesia berdasarkan ringkasan sinopsis film Indonesia Indonesia berdasarkan ringkasan sinopsis film Indonesia, algoritma klasifikasi Support Vector Machines dengan ekstraksi TF-IDF mampu menemukan nilai akurasi dan akurasi dan F1 score terbaik setelah data training (45%).	Pada penelitian ini, masih terdapat kekurangan dan kesalahan pada hasil klasifikasi genre film saat melakukan input ringkasan sinopsis. Hal ini disebabkan karena kurangnya dataset terutama melihat jumlah film Indonesia dengan ringkasan sinopsis dan genre film yang masih dapat digunakan dalam dataset. Perlunya penambahan dataset dan pemilihan data film Indonesia dengan kriteria bahasa Indonesia yang baik dan benar serta menyesuaikan genre film yang ada.	Penelitian tersebut menggunakan synopsis film sebagai objek penelitian, penekanan penelitian ini adalah perbandingan algoritma klasifikasi yang diugunakan untuk mengklasifikasi genre film indonesia.

Tabel 2.1. Lanjutan Matriks literatur review dan posisi penelitian

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
3	Sentiment Analysis about Product and Service Evaluation of PT Telekomunikasi Indonesia Tbk from Tweets Using TextBlob, Naive Bayes & K-NN Method	R. Hermansyah and R. Sarno, *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Semarang, Indonesia, 2020, pp. 511-516	Melakukan analisis sentimen menggunakan algoritma NB, KNN dan Textblob dan mebandingkan hasil pengujiannya untuk menentukan performa terbaik dari ketiga algoritma tersebut.	Dari hasil percobaan, terlihat bahwa klasifikasi K-NN memiliki akurasi yang lebih baik dibandingkan yang lain. Hasil yang diperoleh menunjukkan akurasi dari TextBlob, Naive Bayes dan K-NN sebesar 54.67%, 69.44%, dan 75%. Metode Textblob memiliki performa yang lebih buruk dibandingkan Naive Bayes dan K-NN dalam proses recall karena tipe pengklasifikasi yang tidak supervised type classifier dibandingkan dengan machine learning sebesar 13,68%, 82,76%, dan 56,32%. Akan tetapi dalam proses recall Namun dalam proses recall	Peningkatan lebih lanjut dapat dilakukan dengan menggunakan dataset lebih besar dan lebih kompleks untuk meningkatkan nilai akurasi dalam analisis sentimen. Penelitian di masa depan juga dapat memasukkan setiap aspek dari produk untuk meningkatkan hipotesis sentimen.	Penelitian tersebut menggunakan sentimen yang didapat dari twitter sebagai objek penelitian, penekanan penelitian tersebut adalah perbandingan algoritma klasifikasi yang diugunakan untuk mengklasifikasi sentimen yang merupakan sebuah ulasan untuk Perusahaan Telekomunikasi dalam media sosial twitter.

Tabel 2.1. Lanjutan Matriks literatur review dan posisi penelitian

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
				memberikan keuntungan karena Textblob memaksa untuk menghitung seluruh dataset sebagai data uji memberikan hasil presisi sebesar TextBlob, Naive Bayes, dan K-NN sebesar 94.12%, 64.29%, dan 87.50%. Ketika kami membandingkan setiap pengklasifikasi, metode K-NN memberikan prediksi yang lebih akurat, sehingga dapat disimpulkan bahwa K-NN berkinerja lebih baik daripada Naive Bayes dan Textblob. METODE K-NN berkinerja lebih baik karena kemampuannya untuk menemukan kesamaan antara pengamatan dan sifat yang melekat untuk		

Tabel 2.1. Lanjutan Matriks literatur review dan posisi penelitian

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
				mengoptimalkan secara lokal.		
4	Klasifikasi Teks dengan Naïve Bayes Classifier (NBC) untuk Pengelompokan Keterangan Laporan dan Durasi Recovery Time Laporan Gangguan Listrik PT. PLN (Persero) WS2JB Area Palembang	Eka Afrianti, Fathoni, dan Rahmat Izwan Heroza, JSI : Jurnal Sistem Informasi (E-Journal), VOL.12, NO.1, April 2020	Melakukan pengujian algoritma NBC untuk klasifikasi pengelompokan laporan gangguan listrik	Berdasarkan penelitian yang telah dilakukan dengan menggunakan algoritma Naive Bayes Classifier dalam text mining diketahui bawah laporan yang teklasifikasi dibawah 180 menit ada 64,4% , 34,8% untuk laporan yang terklasifikasi kurang dari 1 hari dan 0,08% untuk laporan yang terklasifikasi lebih dari 1 hari.	Perlunya penelitian terkait seberapa besar pengaruh akurasi terhadap Penggunaan stopwords dan stemming dalam preprocessing.	Perbedaan objek yang diklasifikasikan, penekanan penelitian tersebut adalah menggunakan NBC untuk megklasifikasikan laporan.
5	Implementasi Metode Naive Bayes Classifier (NBC) Pada Komentar Warga Sekolah Mengenai Pelaksanaan Pembelajaran Jarak Jauh (PJJ)	Naomi Chatrina Siregar, Riki Ruli A. Siregar , M. Yoga Distra Sudirman, Jurnal Teknologia, Vol. 3, No. 1, Agustus 2020, ISSN 2654-5683	Melakukan implementasi algoritma NBC untuk klasifikasi Komentar PJJ	Untuk menentukan kategori dilakukan pengujian terhadap 50 data komentar dengan menggunakan algoritma Naïve Bayes Classifier didapatkan bahwa hasil nilai akurasi yaitu sebesar 68%.	Dalam pengembangan selanjutnya, agar dapat melakukan penambahan data training (data latih) dalam jumlah yang lebih banyak dengan tujuan menghasilkan tingkat akurasi yang lebih tinggi karena semakin banyak data latih yang	Penelitian tersebut menggunakan objek yang berbeda yaitu klasifikasi komentar.

Tabel 2.1. Lanjutan Matriks literatur review dan posisi penelitian

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
					digunakan maka akan semakin akurat pula hasil klasifikasi dan dalam pengolahan komentar dapat menggunakan bahasa asing seperti bahasa inggris dan lainnya.	
6	Teks Mining untuk Klasifikasi Keluhan Masyarakat Menggunakan Algoritma Naive Bayes	Dyah Ariyanti, Kurnia Iswardani. IKRAITH-INFORMATIKA Vol 4 No 3 Bulan November 2020	Menguji algoritma NBC untuk klasifikasi keluhan masyarakat	Pengolahan data keluhan masyarakat ini melalui beberapa tahapan teks mining yaitu token, filter, stemming dan analyzing. Setelah melalui tahapan praproses, data tersebut akan dilakukan klasifikasi menggunakan algoritma Naïve Bayes, hasil perhitungan tersebut yang nantinya akan menunjukkan hasil kelas dari setiap data keluhan masyarakat yang masuk baik melalui telepon,	Tidak adanya seleksi fitur	Penelitian tersebut menggunakan objek yang berbeda yaitu klasifikasi keluhan masyarakat.

Tabel 2.1. Lanjutan Matriks literatur review dan posisi penelitian

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
				sms. Penelitian ini menghasikan tingkat akurasi mencapai 95%, schingga dapat mengklasifikasikan keluhan masyarakat tiap-tiap instansi di pemerintah Kota Probolinggo.		

2.3. Landasan Teori

2.3.1. Anime

Anime adalah jenis animasi yang berasal dari Jepang yang dikenal dengan ilustrasi yang terang dan berwarna-warni, menggambarkan karakter yang hidup dalam cerita yang sarat dengan aksi, seringkali dengan tema fantasi atau futuristik. Istilah "anime" umumnya merujuk pada segala jenis animasi yang diproduksi di Jepang (Aziz & Ong, 2023). Anime biasanya memiliki ciri-ciri lewat gambar warna-warni yang menampilkan tokoh-tokoh. Tokoh - tokoh tersebut memiliki keragaman karakter, mulai dari tokoh yang berperan antagonis, protagonis, hingga figuran. Anime sangat dipengaruhi gaya gambar manga yaitu komik khas Jepang (Toi, 2020).

Anime sebagai film animasi juga memiliki banyak Genre, Genre merupakan istilah yang digunakan untuk mengklasifikasikan teks media ke dalam kelompok-kelompok dengan karakteristik sejenis (Rayner et al., 2004) dan Genre merupakan konsep yang diterapkan dalam studi dan teori film untuk mengidentifikasi kesamaan antara kelompok-kelompok film berdasarkan elemen estetika, sosial, institusional, kultural, dan psikologis yang lebih luas. Genre film memperlihatkan keseragaman dalam gaya dan bentuk, tema, serta fungsi komunikatifnya. Oleh karena itu, suatu genre film dibentuk oleh serangkaian konvensi yang memengaruhi produksi karya dalam genre tersebut, serta harapan dan pengalaman penonton. Genre digunakan dalam industri film untuk produksi dan pemasaran, oleh analis dan kritikus untuk menganalisis sejarah film, dan sebagai kerangka kerja bagi penonton dalam memilih dan menikmati film (Bondebjerg, 2015).

Perbedaan antara penceritaan animasi barat (Kartun Amerika) dan timur (Anime Jepang) terletak pada asal-usul cerita, yang dilihat secara berbeda dari sudut pandang barat atau timur. Elemen penceritaan karakter adalah salah satu perbedaan paling signifikan antara penceritaan animasi barat dan timur, Narasi Barat lebih menekankan pada satu karakter sentral (character arc), yang berfungsi sebagai titik fokus narasi. Sepanjang cerita, sang protagonis akan menghadapi berbagai kendala, yang akan ia atasi dengan tekad dan ketabahan untuk mencapai tujuan cerita. Sebaliknya, pendekatan khas pengembangan karakter dalam penceritaan animasi timur adalah salah satu karakter pada akhirnya menjadi protagonis. Alternatifnya, pendekatan barat, di mana tokoh protagonis diperkenalkan di awal cerita, tokoh protagonis bisa diperkenalkan di tengah cerita (Shah et al., 2023). Sebaliknya, dari sudut pandang model penceritaan dan struktur cerita dalam penceritaan animasi barat dan timur, pendekatan yang berbeda terlihat digunakan sebagai bagian dari proses pengembangan cerita. Sebagian besar model bercerita yang tersedia dikembangkan oleh orang-orang Barat dari perspektif ideologi Barat. Mayoritas animasi Amerika menggunakan model dan struktur penceritaan untuk mengembangkan cerita dan plot dengan narasi yang lugas. Sebaliknya, penceritaan animasi timur, khususnya animasi Jepang, sedikit menggunakan model penceritaan atau struktur cerita. Hal ini karena mayoritas cerita Jepang lebih rumit dan kompleks. Banyak anime Jepang tidak mengikuti model penceritaan yang sudah mapan, karena sebagian besar model tersebut dibuat oleh orang Barat. Sebagian besar cerita memiliki pengaruh kuat terhadap budaya dan seni Jepang (Shah et al., 2023).

Sinopsis merupakan suatu ringkasan dari suatu karya atau ide atau gagasan yang ditulis dengan suatu bentuk narasi yang dapat digunakan sebagai suatu parameter untuk membantu penonton dalam mengetahui garis besar alur cerita (Laili et al., 2019).

Sinopsis anime, film live action, dan manga biasanya memiliki inti cerita yang sama, tetapi seringkali ada perbedaan dalam detail, penokohan, dan cara penyampaian cerita. Adaptasi dari anime ke live action dapat mengubah elemen-elemen tertentu untuk menyesuaikan dengan format baru, memperbarui konteks budaya, atau membuat cerita lebih cocok dengan gaya visual yang berbeda. Misalnya, perubahan fokus sutradara terhadap alur pada film yang memengaruhi dominansi tokoh utama dalam film, pemotongan latar yang tidak memiliki fungsi hingga keberadaan subplot tokoh lain (Jayanti, 2020). Selain itu, perbedaan gaya visual dan produksi bisa memengaruhi bagaimana cerita disampaikan.

2.3.2. Naive Bayes

Naive Bayes Classifier (NBC) adalah sebuah teknik prediksi berbasis probabilistik sederhana berdasarkan teorema Bayes teorema Bayes (atau Bayes) dengan asumsi independensi yang kuat (naif) (Prasetyo, 2014), Dalam Bayes (khususnya Naive Bayes), maksud dari independensi yang kuat adalah ketika sebuah fitur data tidak terkait dengan ada atau tidaknya fitur lain dalam data yang sama. Kaitan antara Naive Bayes dengan klasifikasi, korelasi hipotesis, dan bukti klasifikasi adalah hipotesis dalam teorema Bayes merupakan label kelas yang menjadi target dalam pemetaan klasifikasi. Sedangkan bukti adalah fitur yang menjadi input dalam model klasifikasi. Keuntungan menggunakan Naive Bayes

adalah hanya membutuhkan sedikit data pelatihan (Training Data) untuk menentukan estimasi parameter yang dibutuhkan dalam proses klasifikasi (Firmanda & Fitriati, 2018). Naive Bayes sering kali bekerja jauh lebih baik di sebagian besar situasi dunia nyata yang sebanding dengan yang diharapkan (Saleh, 2015).

Model multinomial Naive Bayes memperhitungkan frekuensi setiap kata yang muncul pada dokumen. Misal terdapat dokumen d dan himpunan kelas c. Untuk memperhitungkan kelas dari dokumen d, maka dapat dihitung dengan rumus (Rahman et al., 2017) :

$$P(c|term\ dokumen\ d) = P(c)\ x\ P(\ t_1|c)\ x\ P(\ t_2|c)\ x\ P(\ t_3|c)\ x\ \dots x\ P(\ t_n|c) \quad (1)$$

Keterangan :

P(c) = Probabilitas prior dari kelas c

t_n = Kata dokumen d ke-n

P(c | term dokumen d) = Probabilitas suatu dokumen termasuk kelas c

P(t_n | c) = Probabilitas kata ke-n dengan diketahui kelas c

Probabilitas prior kelas c ditentukan dengan rumus:

$$P(c) = \frac{N_c}{N}$$

Keterangan :

N_c= Jumlah kelas c pada seluruh dokumen

N = Jumlah seluruh dokumen

Sementara rumus Multinomial yang digunakan dengan pembobotan kata TF-IDF adalah sebagai berikut (Rahman et al., 2017) :

$$P(t_n|c) = \frac{W_{ct}+1}{(\sum W' \in VW'_{ct})+B'}$$

Keterangan :

W_ct = Nilai pembobotan tfidf atau W dari term t di kategori c

∑W'€VW' = Jumlah total W dari keseluruhan term yang berada di kategori c.

B' = Jumlah W kata unik (nilai idf tidak dikali dengan tf) pada seluruh dokumen.

2.3.3. Preprocessing

Preprocessing merupakan tahapan awal dari proses text mining. Proses ini mengolah data yang semula 'kotor' menjadi lebih 'bersih' dan terstruktur untuk dapat dilakukan klasifikasi dengan efisien dan menghasilkan hasil uji yang lebih baik.

Case folding merupakan tahapan awal pada Preprocessing yang bertujuan untuk mengubah setiap bentuk kata menjadi lowercase atau huruf kecil.

Data Cleansing adalah proses pembersihan teks dengan menghilangkan data yang tidak relevan seperti username, hahstag, URL, dan Emoticon.

Stopword merupakan daftar kata umum yang tidak memiliki arti penting dan tidak digunakan. Pada proses ini kata umum akan dihapus untuk mengurangi jumlah kata yang disimpan oleh sistem (Manning et al., 2008)

Stemming merupakan proses untuk mencari stem (kata dasar) dari kata hasil stopword removal (filtering). Terdapat dua aturan dalam melakukan stemming yaitu dengan pendekatan kamus dan pendekatan aturan (Utomo & Oktora, 2017).

Tokenisasi adalah proses untuk memotong kata dari teks menjadi beberapa token. Pada proses ini akan menghilangkan spasi.

2.3.4. TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) adalah metode yang digunakan untuk menghitung bobotan setiap kata yang telah diekstrak. Penggunaan metode ini umumnya dilakukan untuk mengihitung kata umum yang ada pada information retrieval. Model pembobotan TF-IDF merupakan metode yang mengintegrasikan model term frequency (tf) dan inverse document frequency (idf). Term frequency (tf) merupakan proses untuk menghitung jumlah kemunculan term dalam satu dokumen dan inverse document frequency (idf) digunakan untuk menghitung term yang muncul di berbagai dokumen (komentar) yang dianggap sebagai term umum, yang dinilai tidak penting (Manning et al., 2008).

Tahapan pembobotan dengan TF-IDF adalah:

a. Hitung *term frequency* ($tf_{t,d}$)
b. Hitung *weighting term frequency* (W_{tf})

$$W_{tf} = \begin{cases} 1 + log10tf_{t,d}, jika\ tf_{t,d} > 0 \\ 0, \end{cases} \quad (2)$$

c. Hitung *document frequency* (df)
d. Hitung bobot *inverse document frequency* (idf)

$$idf_t = Log\frac{N}{df_t} \quad (3)$$

e. Hitung nilai bobot TF-IDF

$$W_{t,d} = Wtf_{t,d}\ x\ idf_t \quad (4)$$

Keterangan :

$tf_{t,d}$ = frekuensi *term*

$Wtf_{t,d}$ = bobot frekuensi *term*

df = jumlah frekuensi dokumen yang mengandung *term*

N = jumlah total dokumen

$W_{t,d}$ = bobot TF-IDF

2.3.5. Mutual Infomation

Mutual Information (MI) adalah pengukuran jumlah informasi yang dikandung oleh satu variabel acak tentang variabel acak lainnya. MI adalah pengurangan ketidakpastian suatu variabel acak yang disebabkan oleh informasi dari variabel acak lainnya (Cover & Thomas, 2005). MI menentukan korelasi antara dua kata dalam kumpulan data, jika nilai MI besar maka kedua istilah tersebut sering muncul bersama sehingga berhubungan secara semantik. Sebaliknya, nilai MI yang kecil berarti bahwa ketika salah satu dari mereka muncul maka yang lain tidak muncul, yang mengindikasikan tidak ada hubungan semantik. Rumus untuk menentukan MI adalah sebagai berikut :

$$I(x,y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \quad (5)$$

Dengan p(x,y) sebagai probabilitas gabungan dari x dan y, p(x) sebagai probabilitas x, dan p(y) sebagai probabilitas y.

2.3.6. Confusion Matrix

Confusion Matrix atau matriks kebingungan adalah tabel yang sering digunakan untuk menggambarkan kinerja model klasifikasi pada serangkaian data uji yang nilai sebenarnya diketahui. Tabel berikut menunjukkan Confusion Matrix untuk model klasifikasi dua kelas (Herlambang et al., 2015).

Tabel 2.2. Confusion Matrix

	Hasil Prediksi: Salah	Hasil Prediksi: *Benar*
Asli: Salah	TN	FP
Asli: *Benar*	FN	TP

Pada penelitian ini, arti entri pada *confusion matrix* adalah sebagai berikut:

- TP untuk prediksi benar bahwa data merupakan benar
- TN untuk prediksi benar bahwa data merupakan salah
- FP untuk prediksi salah bahwa data merupakan benar
- FN untuk prediksi salah bahwa data merupakan salah

Rumus untuk menghitung kinerja sistem menggunakan entri dari *confusion matrix* adalah sebagai berikut :

True Positive Rate = $\frac{TP}{TP+FN}$

True Negative Rate = $\frac{TN}{TN+FP}$

Akurasi = $\frac{TP+TN}{TP+TN+FP+FN} \times 100\%$

Presisi = $\frac{TP}{TN+FN} \times 100\%$

Recall = $\frac{TP}{TP+FN} \times 100\%$

F-Measure = $\frac{2}{\frac{1}{Presisi}+\frac{1}{Akurasi}}$

Precision adalah proporsi kasus yang diprediksi positif yang juga positif benar pada data sebenarnya. Recall adalah proporsi dari kasus positif kejadian sebenarnya yang diprediksi positif benar. (Powers, 2020)

2.3.7. K-Nearest Neighbors

Salah satu metode klasifikasi paling sederhana yang digunakan dalam penambangan data dan pembelajaran mesin adalah K-Nearest Neighbor (KNN). Ini adalah metode klasifikasi yang paling diterima karena kemudahan dan efisiensi praktisnya: itu tidak mengharuskan pemasangan model dan telah terbukti memiliki kinerja yang unggul untuk mengklasifikasikan beberapa jenis data. Namun, kinerja klasifikasi superior KNN sangat tergantung pada metrik yang digunakan untuk menghitung jarak berpasangan antara titik data. Aturan klasifikasi KNN dibuat oleh sampel pelatihan saja, tanpa data tambahan lainnya. Dalam pendekatan yang lebih rumit, klasifikasi KNN, menemukan sekelompok objek k dalam set pelatihan yang paling dekat dengan objek tes, dan mendasarkan penugasan label pada dominasi kelas tertentu di lingkungan ini. Algoritma K-Nearest Neighbor (KNN) adalah metode untuk mengklasifikasikan objek berdasarkan contoh pelatihan terdekat di ruang fitur. KNN adalah jenis pembelajaran berbasis contoh, atau pembelajaran malas di mana fungsinya hanya didekati secara lokal dan semua perhitungan ditangguhkan hingga klasifikasi. (Nikhath et al., 2016)

Pada metode KNN memerlukan perhitungan jarak untuk mengetahui kedekatan dari setiap dokumen, untuk perhitungan jaraknya digunakan metode yaitu *Eucledian Distance*. Kedekatan dokumen dihitung berdasarkan jarak *eucledian* terdekat. Berikut adalah tahapan dari metode KNN yang akan dijalankan:

a. Mengubah data dokumen yang telah memiliki bobot ke dalam bentuk vektor fitur.
b. Menentukan nilai k

c. Menghitung jarak antara data uji dengan data training
d. Mengurutkan jarak berdasarkan jarak terdekat
e. Mengambil tetangga sejumlah k
f. Menghitung tetangga dengan dengan kategori yang sama dan dicari yang terbanyak
g. Mendapatkan hasil prediksi

BAB III

METODE PENELITIAN

3.1. Jenis, Sifat, dan Pendekatan Penelitian

3.1.1. Jenis Penelitian

Jenis penelitian yang digunakan dalam penelitian ini merupakan penelitian kuantitatif, yakni penelitian dengan melakukan perhitungan matematis sehingga menemukan hasil yang diinginkan.

3.1.2. Sifat Penelitian

Sifat penelitian yang dilakukan merupakan eksperimental yang dalam penelitian ini dilakukan pengujian klasifikasi teks dengan Naive Bayes dan membandingkan akurasi, presisi, recall, dan f-measure pada masing-masing kombinasi fitur yang digunakan saat preprocessing dalam melakukan klasifikasi genre berdasarkan sinopsis anime.

3.1.3. Pendekatan Penelitian

Pendekatan dalam penelitian ini ialah pendekatan kuantitatif dimana penelitian yang akan dilakukan sesuai dengan alur yang telah dibuat oleh peneliti.

3.2. Metode Pengumpulan Data

Pengujian menggunakan 1000 data sinopsis anime dengan genre sebanyak 4 jenis yaitu Fantasy, Mystery, Romance dan Sports yang masing-masing genre memiliki 250 data sinopsis. Data didapatkan dengan cara melakukan scraping pada situs MyAnimeList lalu masing-masing sinopsis dan genrenya akan disatukan

menjadi sebuah dokumen .csv. Dalam penelitian yang dilakukan data yang digunakan hanyalah data sinopsis anime yang menggunakan Bahasa Inggris.

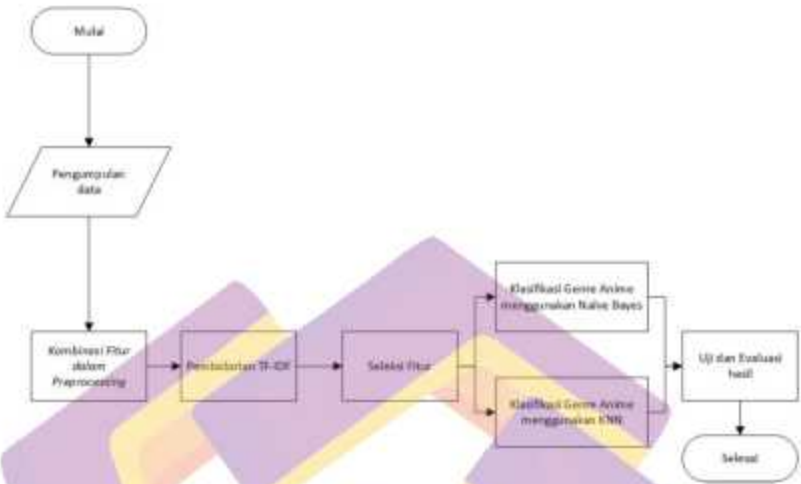
3.3. Metode Analisis Data

Sebelum melakukan analisis data tahap awal yang dilakukan adalah melakukan proses pembersihan dataset sehingga dapat digunakan dalam tahap selanjutnya.

Selanjutnya melakukan pembagian data menjadi data testing dan data training lalu klasifikasi data menggunakan algoritma Naive Bayes dan KNN.

3.4. Alur Penelitian

Pada gambar 3.14 digambarkan secara umum bagaimana penelitian yang peneliti lakukan, yaitu dimulai dari pengumpulan data dilanjutkan dengan menginputkan data ke sistem yang selanjutnya dilakukan tahap preprocessing, Setelah preprocessing dilakukan, dilakukan terlebih dahulu pembobotan TF-IDF dan seleksi fitur, setelah itu data akan diklasifikasikan menggunakan algoritma Naive Bayes dan KNN agar menghasilkan output berupa genre yang tepat. Terakhir akan dilakukan evaluasi dari semua hasil yang didapatkan.



Gambar 3.1. Flowchart Alur Penelitian

3.4.1. Pengumpulan Data

Pada tahap ini dilakukan scraping data dari website MyAnimeList (MAL) untuk mengumpulkan data anime seperti genre dan sinopsis anime. Genre yang dikumpulkan terdiri dari 4 genre yaitu Fantasy, Mystery, Romance, dan Sports. Setelah data dikumpulkan selajutnya data ke dalam file csv. Pengumpulan dataset terdiri dari beberapa tahapan yakni scraping, data cleaning, pemisahan data, dan penggabungan data.

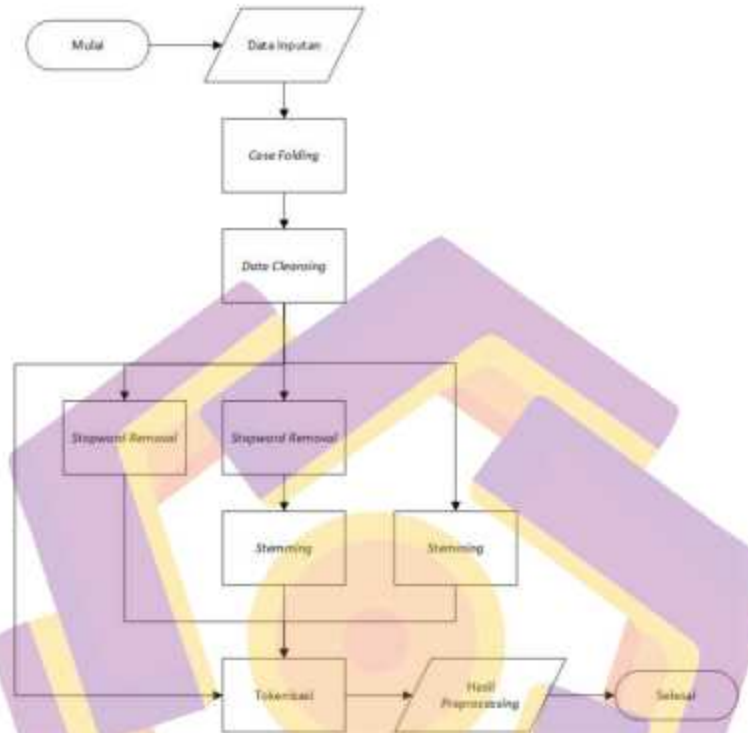
3.4.2. Preprocessing

Tahap preprocessing ini terdiri atas tahap case folding, data cleansing, stopword removal, stemming, dan tokenisasi. Penelitian ini membahas mengenai teknik preproceesing dengan melakukan beberapa skenario pengujian kombinasi teknik preprocessing untuk mengetahui teknik preprocessing yang menghasilkan matrix penilaian paling optimal serta pengaruhnya terhadap klasifikasi genre anime

berdasarkan sinopsis. Terdapat 4 Kombinasi fitur preprocessing yang berbeda yang akan digunakan pada penelitian ini, yang membedakannya hanya penggunakan fitur Stopword Removal dan/atau Stemming. Tidak ada proses yang mutlak dalam fitur yang digunakan dalam preprocessing, pada penelitian ini dilakukan perubahan fitur preprocessing dengan tujuan menganalisis dampak sebuah fitur tersebut terhadap hasil klasifikasi yang dilakukan. Stopword removal dan Stemming karena kedua fitur tersebut terjadi perubahan struktur bahasa, Dalam beberapa kasus, Stopword Removal dapat menyebabkan hilangnya banyak informasi dan dapat terjadi perubahan makna pada beberapa kata setelah stemming dilakukan. Perbedaan masing-masing kombinasi fitur yang digunakan dalam penelitian ini dapat terlihat pada tabel 3.1. Gambar 3.2 menampilkan Alur preprocessing yang terdiri atas tahap case folding, data cleansing, normalisasi bahasa, stopword removal, stemming, dan tokenisasi.

Tabel 3.1. Kombinasi Fitur

	Kombinasi 1	Kombinasi 2	Kombinasi 3	Kombinasi 4
Stopword Removal	Tidak	Ya	Tidak	Ya
Stemming	Tidak	Tidak	Ya	Ya



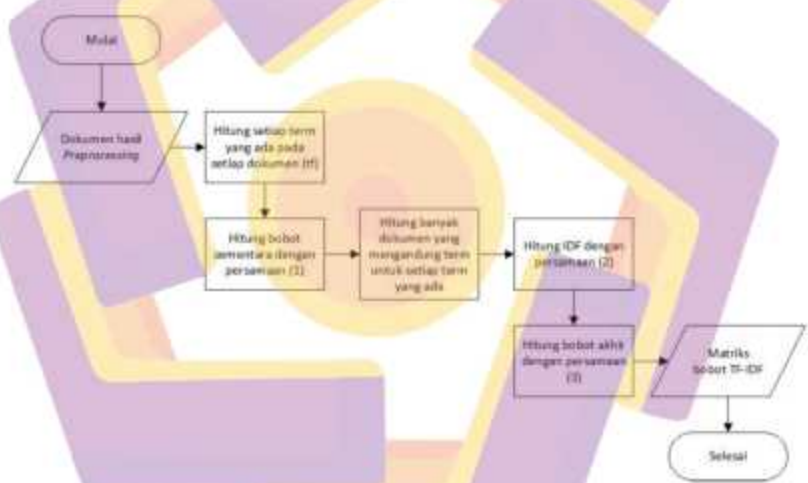
Gambar 3.2. Alur Preprocessing

Proses awal dari preprocessing yaitu Proses case folding yaitu secara garis besar hanya mengubah seluruh data sinopsis menjadi huruf kecil dan menyimpannya kembali. Proses stopword removal digunakan untuk menghilangkan kata-kata yang tidak penting dan sering muncul seperti "he, she, is, we, our, to, be, him, his". Proses data cleansing yakni menghapus semua karakter yang tidak relevan seperti username, hahstag, URL, dan Emoticon. Proses stemming merupakan proses mengubah kata menjadi kata dasar seperti "looking" menjadi "look". Proses terakhir dari Preprocessing adalah tokenization,

tokenization yaitu memecah teks menjadi bentuk token dengan contoh seperti teks "we live in a society" menjadi ['we', 'live', 'in', 'a', 'society'].

3.4.3. Pembobotan TF-IDF

Data yang sudah melalui tahapan preprocessing akan ditentukan bobot untuk setiap term yang ada. Pembobotan term dilakukan dengan menghitung TF-IDF berdasarkan persamaan (2),(3), dan (4) untuk di setiap term yang ada. Hasil dari bobot TF-IDF ini akan digunakan pada klasifikasi dengan Naive Bayes dan KNN.



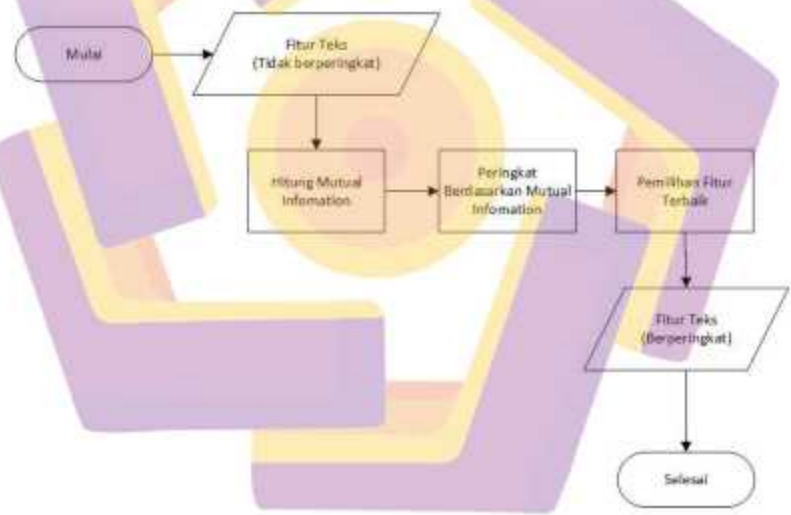
Gambar 3.8. Alur TF-IDF

Terkait Gambar 3.8, pembobotan term dengan TF-IDF dimulai dengan menghitung term frequency (tf) atau frekuensi kemunculan suatu term pada satu email. Setelah itu dilakukan perhitungan bobot sementara dengan persamaan (2) dan didapat Wtf. Untuk menurunkan nilai term yang sering muncul, dihitung document frequency (df) atau jumlah dokumen yang berisi suatu term, yang

dilanjutkan dengan perhitungan inverse-document frequency (idf) dengan persamaan (3). Untuk mendapatkan bobot tf-idf, dilakukan perhitungan sesuai dengan persamaan (4) dan didapatkan hasil berupa matriks bobot term. Hasil pembobotan TF-IDF akan disimpan ke dalam bentuk Matrix TF-IDF.

3.4.4. Seleksi Fitur

Pada tahap ini, data matriks TF-IDF akan dilakukan seleksi fitur menggunakan metode Mutual Information untuk mengurangi fitur yang tidak cukup relevan. Gambar 3.9 merupakan alur seleksi fitur yang diterapkan pada penelitian ini.



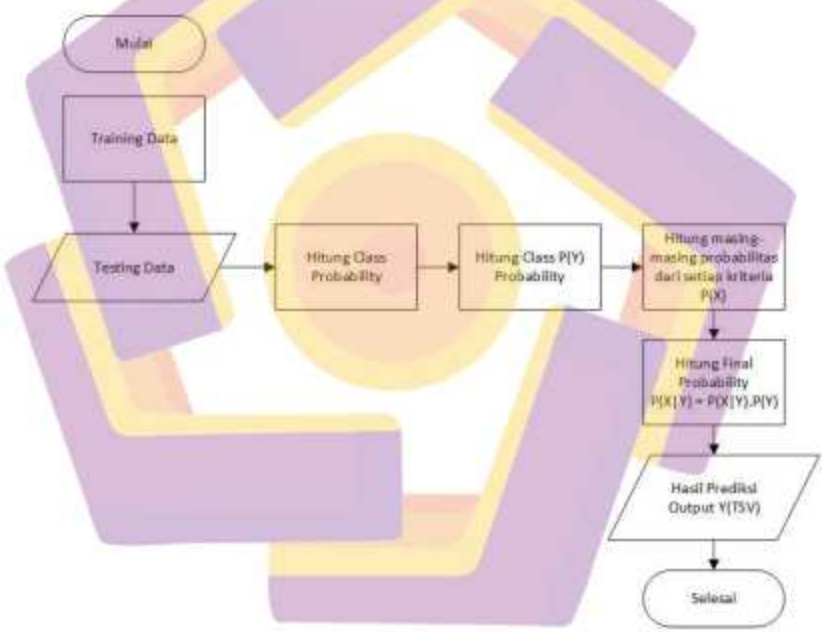
Gambar 3.9. Alur Seleksi Fitur

Merujuk pada Gambar 3.9., secara garis besar alur seleksi fitur akan menghitung nilai mutual information kemudian fitur-fitur tersebut akan diurutkan berdasarkan peringkat nilai MI dan hanya fitur terbaik berdasarkan nilai MI yang

akan dipilih(disimpan) dan fitur yang tidak dipilih akan terhapus tidak digunakan dalam tahap klasifikasi.

3.4.5. Klasifikasi Naive Bayes

Pada tahap ini, Klasifikasi Genre dilakukan menggunakan data matriks TF-IDF yang telah dilakukan seleksi fitur menggunakan metode Mutual Information. Proses Klasifikasi ini dilakukan menggunakan algoritma Naive Bayes. Gambar 3.10 merupakan alur Klasifikasi Naive Bayes yang diterapkan pada penelitian ini.

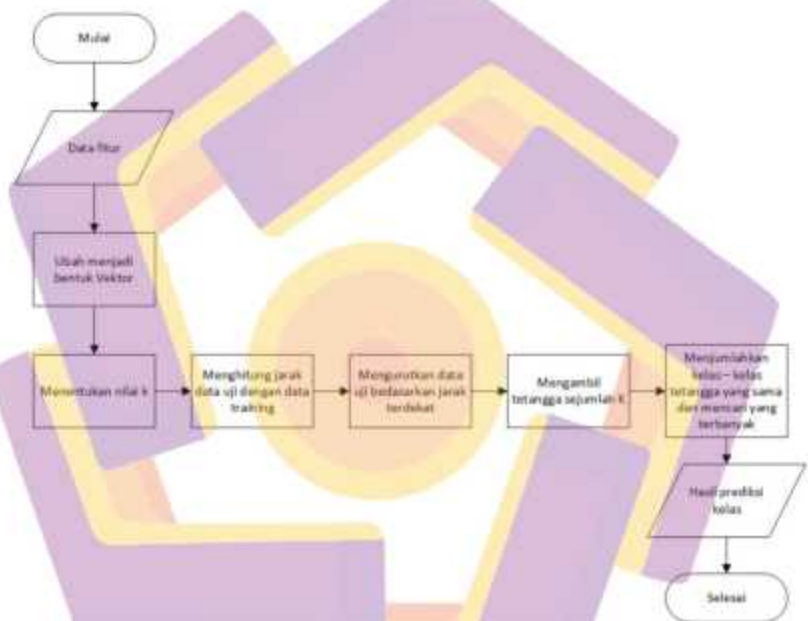


Gambar 3.10. Alur Klasifikasi Naive Bayes

Merujuk pada Gambar 3.10., secara garis besar alur klasifikasi menggunakan naive bayes akan menghitung Probabilitas dari data testing terhadap probabilitas dari setiap kriteria untuk mendapatkan hasil prediksi kelas yang benar.

3.4.6. Klasifikasi K-Nearest Neighbors

Pada tahap ini, Klasifikasi Genre dilakukan menggunakan data matriks TF-IDF yang telah dilakukan seleksi fitur menggunakan metode Mutual Information. Proses Klasifikasi ini dilakukan menggunakan algoritma KNN. Gambar 3.11 merupakan alur seleksi fitur yang diterapkan pada penelitian ini.



Gambar 3.11. Alur Klasifikasi KNN

Merujuk pada Gambar 3.11., secara garis besar alur klasifikasi menggunakan KNN yaitu menentukan nilai k terlebih dahulu lalu akan menghitung dan menyortir jarak data testing terhadap data training dari yang tedekat dan diambilnya tetangga sejumlah nilai k untuk menentukan prediksi kelas untuk data uji tersebut.

3.4.7. Uji dan Evaluasi Hasil

Hasil pengujian Klasifikasi yang dilakukan menggunakan Naive Bayes dan KNN akan dievaluasi menggunakan confusion matrix dan mendapatkan hasil berupa Accuracy, Precision, Recall, dan F-Measure.

BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

4.1. Pengumpulan Data

Pengumpulan dataset terdiri dari beberapa tahapan yakni scraping, data cleaning, pemisahan data, dan penggabungan data. Alur proses pengumpulan dataset ditunjukkan pada gambar 4.1

Pengumpulan dataset dilakukan pada tanggal 5 November 2023. Pengumpulan dataset dilakukan dengan cara melakukan scraping pada situs MyAnimeList (MAL) menggunakan bahasa pemrograman Python.

Gambar 4.1. Alur Pengumpulan Data

4.1.1. Scraping

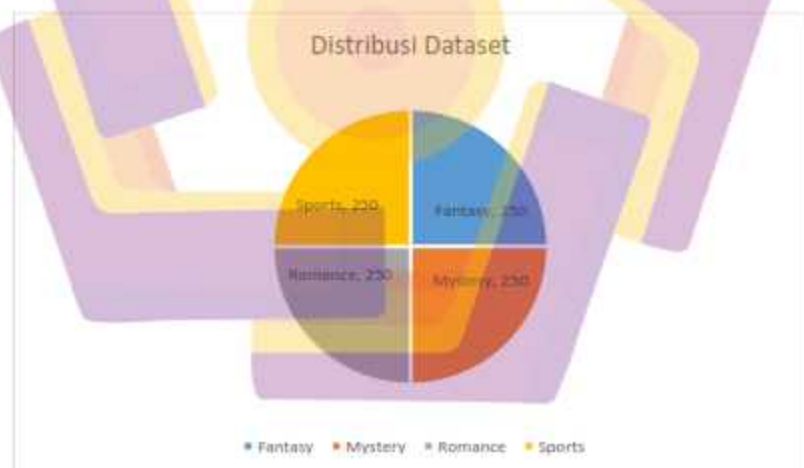
Pada tahap ini dilakukan scraping data dari website MyAnimeList (MAL) untuk mengumpulkan data anime seperti genre dan sinopsis anime. Setelah data dikumpulkan selajutnya data disimpan ke dalam file csv.

Genre yang digunakan dari penelitian ini adalah :

1. Genre Fantasy, dalam anime seringkali melibatkan dunia atau elemen supernatural, magis, atau mitologis. Dunia dalam anime fantasy biasanya memiliki aturan yang berbeda dari dunia nyata, seperti keberadaan sihir, monster, atau makhluk mitos. Contoh anime yang termasuk dalam genre ini adalah "Re:Zero kara Hajimeru Isekai Seikatsu," "Kono Subarashii Sekai ni Shukufuku wo!," dan "Gate: Jieitai Kanochi nite, Kaku Tatakaeri"
2. Genre Mystery, berfokus pada teka-teki atau misteri yang harus dipecahkan. Cerita dalam anime mystery sering kali melibatkan detektif, penyelidikan, atau situasi yang penuh ketegangan dan kejutan. Contoh anime dalam genre ini termasuk "Detective Conan," "GOSICK," dan "Persona 4 the Animation"
3. Genre Romance, berfokus pada hubungan percintaan antara karakter utama. Cerita sering kali melibatkan perkembangan perasaan cinta, tantangan dalam hubungan, dan momen-momen emosional yang intens. Beberapa contoh anime romance adalah "Chuunibyou demo Koi ga Shitai!," "Nisekoi," dan "Sakura-sou no Pet na Kanojo"

4. Genre Sports, mengangkat tema tentang olahraga dan kompetisi. Fokus cerita biasanya pada karakter yang berusaha menjadi atlet terbaik, pertandingan, dan kerja sama tim. Beberapa contoh anime sports adalah "Captain Tsubasa," "Kuroko's Basketball," dan "Eyeshield 21."

Data yang akan digunakan sebanyak 1000 sinopsis yang masing masing terbagi menjadi 250 Sinopsis untuk genre Fantasy, 250 Sinopsis untuk genre Mystery, 250 Sinopsis untuk genre Romance, dan 250 Sinopsis untuk genre Sports. Dari dataset yang digunakan, data dibagi menjadi proporsi 80% untuk data training dan 20% untuk data testing, proses pembagian data dilakukan secara acak dan setiap label mempunyai jumlah anggota yang sama saat dilakukan pembagian.



Gambar 4.2. Grafik Distribusi Dataset

Seperti yang terlihat pada gambar 4.2., semua data terdistribusi secara sama rata ke dalam 4 kelas genre sebanyak masing-masing 250 data yang menghasilkan

total data menjadi 1000 data, hal ini dilakukan untuk mecegah dataset yang tidak seimbang atau disebut dengan istilah Imbalanced Dataset. Dataset yang tidak seimbang kelemahan salah satunya adalah Model yang dilatih pada dataset tidak seimbang cenderung bias terhadap kelas mayoritas, Akibatnya algoritma klasifikasi yang digunakan mungkin berkinerja baik dalam memprediksi kelas mayoritas tetapi buruk dalam memprediksi kelas minoritas (Ali et al., 2019).

4.1.2. Data Cleaning

File yang sudah disimpan dilakukan tahap pembersihan data yakni proses untuk menghapus data yang terduplikat dan data yang kosong. Proses pembersihan data melibatkan perbaikan atau penghapusan data yang tidak akurat, rusak, tidak sesuai format, duplikat, atau tidak lengkap dalam kumpulan data. Ketika data dari berbagai sumber digabungkan, terdapat banyak kesempatan di mana data dapat menjadi duplikat atau diberi label yang salah. Kehadiran data yang tidak benar dapat mengakibatkan ketidakandalan hasil dan algoritma, meskipun kesannya tampak benar. Tidak ada pendekatan yang bersifat mutlak dalam menentukan langkah-langkah yang sesuai dalam proses pembersihan data karena setiap dataset memiliki karakteristik yang berbeda. Dalam penelitian ini, proses pembersihan data melibatkan penghapusan entri yang duplikat, serta penghapusan sinopsis dan genre yang tidak memiliki isi.

4.1.3. Pemisahan Data

Jumlah genre yang digunakan pada penelitian ini adalah sebanyak empat genre, dari hasil scraping data yang telah didapatkan, Setiap genre dan sinopsisnya dipisah ke dalam empat file csv. Pemisahan data dilakukan menggunakan bahasa

pemrograman Python dengan aturan per skenario misalnya ketika mencari genre 'Fantasy', maka ketika mencari genre 'Fantasy' tersebut harus tidak menyertakan genre 'Mystery', 'Romance', dan 'Sports. Begitu pula saat skenario yang mencari genre lainnya, hal ini dilakukan karena satu sinopsis bisa saja memiliki dua atau lebih genre.

4.1.4. Penggabungan data

Setelah dilakukan pemisahan, empat file csv yang telah dipisah tersebut digabungkan ke dalam satu file csv untuk selanjutnya dilakukan preprocessing. Tabel 4.1. Berikut ini adalah preview dataset sinopsis anime yang telah digabungkan ke dalam satu file csv.

Tabel 4.1. Preview Dataset Sinopsis

Index	Sinopsis	Label
1	When Subaru Natsuki leaves the convenience store, the last thing he expects is to be wrenched from his everyday life and dropped into a fantasy world. Things are not looking good for the bewildered teenager, however, not long after his arrival, he is attacked by some thugs. ...	Fantasy
2	In class 3-3 of Yomiyama North Junior High, transfer student Kouichi Sakakibara makes his return after taking a sick leave for the first month of school. Among his new classmates, he is inexplicably drawn toward Mei Misaki—a reserved girl with an eyepatch whom he met in the hospital during his absence. But none of his classmates acknowledge her existence; ...	Mystery

Tabel 4.1. Lanjutan Preview Dataset Sinopsis

3	Everybody has had that stage in their life where they have thought themselves to be special, different from the masses of ordinary humans. They might go as far as seeing themselves capable of wielding mystical powers, or maybe even believe themselves to have descended from a fantasy realm. ...	Romance
4	Famous racehorses that have left behind worthy legacies, unique as they can be, are reincarnated as horse girls in a parallel world. In this life, they start their journey anew as they continue to race and perhaps relive the success they once lived through. ...	Sports

Analisa sinopsis selanjutnya menggunakan wordcloud bertujuan untuk melihat distribusi kemunculan kata dalam sinopsis. Gambar 4.3 merupakan hasil wordcloud sinopsis yang telah dilakukan scraping hingga penggabungan. Dari wordcloud tersebut dapat dilihat kata-kata yang sering muncul seperti 'year', 'first', 'story', 'day', 'losing', 'school', 'high'. Kata- kata tersebut memiliki ukuran yang lebih besar dibandingkan dengan kata-kata yang lain. Berdasarkan dari worcloud ini, dapat dikatakan bahwa sebagian besar tema atau latar belakang dari sinopsis yang telah digabungkan yaitu berlatar belakang sekolah atau murid sekolah.

Gambar 4.3. Wordcloud dari Dataset

4.2. Data Preprocessing

Tahap text preprocessing yaitu tahap dimana perbaikan teks sinopsis dari teks yang sulit diketahui oleh sistem menjadi mudah untuk diproses menggunakan metode klasifikasi. Tahap preprocessing terdiri atas tahap case folding, data cleansing, stopword removal, stemming, dan tokenisasi. Masing-masing kombinasi akan dilakukan disimpan kedalam file .csv terpisah sesuai dengan penomoran kombinasi fitur yang digunakan, yang membedakannya hanya penggunakan fitur Stopword Removal dan/atau Stemming. Perbedaan masing-masing kombinasi fitur yang digunakan dalam penelitian ini dapat terlihat pada tabel 4.2.

Tabel 4.2. Perbedaan Kombinasi Fitur

	Kombinasi 1	Kombinasi 2	Kombinasi 3	Kombinasi 4
Case Folding	✓	✓	✓	✓
Data Cleansing	✓	✓	✓	✓
Stopword Removal	✗	✓	✗	✓
Stemming	✗	✗	✓	✓
Tokenisasi	✓	✓	✓	✓

4.3.1. Case Folding

Tahap case folding adalah tahap untuk mengubah semua huruf menjadi lower case atau huruf kecil. Tujuan Case Folding ini adalah menghasilkan data yang lebih terstruktur dan penyamarataan penggunaan semua huruf kapital. Tabel 4.3

merupakan contoh hasil dari proses case folding sinopsis paragraf pertama dari anime Re:Zero kara Hajimeru Isekai Seikatsu yang dirilis pada tahun 2016.

Tabel 4.3. Hasil Case Folding

Sebelum Case Folding	Sesudah Case Folding
When Subaru Natsuki leaves the convenience store, the last thing he expects is to be wrenched from his everyday life and dropped into a fantasy world. Things are not looking good for the bewildered teenager; however, not long after his arrival, he is attacked by some thugs.	when subaru natsuki leaves the convenience store, the last thing he expects is to be wrenched from his everyday life and dropped into a fantasy world. things are not looking good for the bewildered teenager; however, not long after his arrival, he is attacked by some thugs.

4.3.2. Data Cleansing

Selanjutnya sistem melalui tahapan cleansing, suatu langkah untuk menghilangkan semua karakter numerik, simbol, tanda baca, tag HTML (Hypertext Markup Language), dan juga URL (Uniform Resource Locator). Proses ini akan menghilangkan tanda baca, dan emoticon. Beberapa data sinopsis terdapat karakter penulisan bahasa Jepang (hiragana (平仮名), katakana (片仮名) dan kanji (漢字)), Karakter penulisan bahasa jepang tersebut juga dihapus dari data sinopsis melalui tahapan ini. Selain itu pula, pada setiap sinopsis tertera keterangan dari mana sinopsis tersebut berasal, keterangan tersebut akan dihapus melalui tahapan ini seperti contohnya "[Written by MAL Rewrite]", "(Source: ANN)", "(Source:

AniDB)", dll karena kata-kata tersebut dapat menjadi fitur yang termasuk ke dalam outliers dan noise yang dapat memperburuk hasil akurasi.

Hasil data cleansing dapat dilihat pada Tabel 4.4 terdapat simbol titik (.), titik koma (;), dan koma (,) dihapus dari dataset sinopsis.

Tabel 4.4. Hasil Data Cleansing

Sebelum Data Cleansing	Sesudah Data Cleansing
when subaru natsuki leaves the convenience store, the last thing he expects is to be wrenched from his everyday life and dropped into a fantasy world. things are not looking good for the bewildered teenager; however, not long after his arrival, he is attacked by some thugs.	when subaru natsuki leaves the convenience store the last thing he expects is to be wrenched from his everyday life and dropped into a fantasy world things are not looking good for the bewildered teenager however not long after his arrival he is attacked by some thugs

4.3.3. Stopwords Removal

Kemudian dilanjutkan ke tahap stopword removal untuk menghilangkan semua kata-kata yang merupakan stopword pada Bahasa Inggris, pada proses ini melibatkan library dari NLTK dan Menggunakan corpus stopwords dari library tersebut. Pada tabel 4.5 terdapat kata-kata seperti 'the', 'he', 'is', 'to', 'be', 'from' ,'his', 'and', 'into' 'a', 'are', 'not', 'for', dan 'are' merupakan kata yang termasuk dalam stopword yang dihapus dari dataset sinopsis. Hasil proses stopword removal dapat dilihat pada Tabel 4.5 sebelum dilakukan stopwords removal dan sesudahnya.

Tabel 4.5. Hasil Stopwords Removal

Sebelum Stopwords Removal	Sesudah Stopwords Removal
when subaru natsuki leaves the convenience store the last thing he expects is to be wrenched from his everyday life and dropped into a fantasy world things are not looking good for the bewildered teenager however not long after his arrival he is attacked by some thugs	when subaru natsuki leaves convenience store last thing expects wrenched everyday life dropped fantasy world things looking good bewildered teenager however long arrival attacked thugs

4.3.4. Stemming

Proses selanjutnya dilanjutkan dengan tahapan stemming dimana pada stemming akan mencari kata dasar dari sebuah kata, menghilangkan semua imbuhan baik yang terdiri dari awalan, sisipan dan akhiran, proses ini akan melibatkan library Porter Stemmer untuk Bahasa Inggris. Terlihat pada tabel 4.6 kata-kata berimbuhan seperti 'leaves', 'convenience', 'expects', 'wrenched', 'dropped', 'fantasy', 'things', 'looking', 'bewildered', 'teenager', 'however', 'arrival', 'attacked', 'thugs'. Hasil proses stemming dapat dilihat pada Tabel 4.6 sebelum dilakukan stemming dan sesudahnya.

Tabel 4.6. Hasil Stemming

Sebelum Stemming	Sesudah Stemming
when subaru natsuki leaves convenience store last thing expects wrenched everyday life dropped fantasy world things looking good bewildered teenager however long arrival attacked thugs	when subaru natsuki leav conveni store last thing expect wrench everyday life drop fantasi world thing look good bewild teenag howev long arriv attack thug

4.3.5. Tokenization

Tokenisasi adalah teknik untuk memecah teks tertentu menjadi fragmen-fragmen kecil atau token. Token bisa berupa kata-kata atau karakter. Hasil tokenisasi diperlihatkan pada tabel 4.7. Terdapat 25 kata berbeda yang telah menjadi token.

Tabel 4.7. Hasil Tokenization

Sebelum Tokenization	Sesudah Tokenization
when subaru natsuki leav conveni store last thing expect wrench everyday life drop fantasi world thing look good bewild teenag howev long arriv attack thug	[when, subaru, natsuki, leav, conveni, store, last, thing, expect, wrench, everyday, life, drop, fantasi, world, thing, look, good, bewild, teenag, howev, long, arriv, attack, thug]

4.3. Pembobotan TF-IDF

Pada tahap ini akan melakukan ekstraksi fitur, fitur yang diekstraksi pada proses ini adalah Term Frequency-Inverse Document Frequency (TF-IDF). Masing-masing kombinasi akan dilakukan pembobotan TF-IDF, peneliti menggunakan library 'tfidfvectorizer' sehingga hasil TF-IDF berupa matriks yang kemudian disimpan kedalam file .csv terpisah sesuai dengan penomoran kombinasi fitur yang digunakan.

4.4. Seleksi Fitur

Setelah dilakukan TF-IDF, data tersebut akan dilakukan seleksi fitur menggunakan Mutual Information, setiap terms yang ada akan dihitung menggunakan rumus (5) dan menghasilkan nilai MI, setiap terms akan diurutkan berdasarkan nilai tersebut dan kemudian terms yang memiliki nilai MI sebesar 0 akan dieliminasi dan tidak akan digunakan dalam tahap klasifikasi. Berikut ini adalah perubahan jumlah fitur sebelum dan sesudah dilakukan seleksi fitur menggunakan Mutual Information :

Tabel 4.8. Perubahan jumlah fitur

	Kombinasi 1	Kombinasi 2	Kombinasi 3	Kombinasi 4
Jumlah fitur sebelum Seleksi Fitur	12947	12885	9474	9417
Jumlah fitur setelah Seleksi Fitur	6516	6462	4730	4702

Gambar 4.3. Grafik Perubahan Jumlah Fitur

Tabel 4.8 dan Gambar 4.3 di atas menunjukkan perubahan jumlah fitur sebelum dan setelah proses seleksi fitur pada empat kombinasi yang berbeda. Setiap kombinasi merupakan konfigurasi text preprocessing yang berbeda dari dataset. Dari Grafik tersebut, dapat disimpulkan bahwa seleksi fitur menghasilkan pengurangan jumlah fitur yang signifikan dalam setiap kombinasi, sehingga memungkinkan untuk mengurangi dimensi data dan mungkin meningkatkan kinerja klasifikasi atau model yang akan digunakan.

4.5. Klasifikasi Menggunakan K-Nearest Neighbors

Proses selanjutnya setelah melakukan proses preprocessing dan seleksi fitur yang dibutuhkan adalah proses klasifikasi menggunakan algoritma KNN. Tahap ini akan menerapkan algoritma KNN yang telah dipaparkan pada sub-bab sebelumnya.

Tidak ada nilai mutlak untuk nilai k pada algoritma K-Nearest Neighbors (KNN) karena nilai k yang optimal sangat tergantung pada karakteristik dataset

yang digunakan, termasuk distribusi data, ukuran dataset, dan kompleksitas masalah.

Pada penelitian ini, nilai k didapatkan dengan cara mengikuti suatu pendekatan, yaitu aturan empiris di mana nilai k yang sama dengan akar kuadrat dari jumlah sampel, yang biasanya menghasilkan hasil yang lebih akurat (Sun et al., 2018). Park dan Lee (2020) menyatakan bahwa aturan praktis empiris yang baik adalah menetapkan k sebagai akar kuadrat dari ukuran data. Ketika memiliki data uji baru dan ingin mengetahui k tetangga terdekat yang paling dekat dan yang paling "mirip" dengannya, Angka k biasanya dipilih sebagai akar kuadrat dari n, jumlah total titik dalam set data pelatihan. (Dengan demikian, jika n adalah 400, k = 20) (Nadkarni, 2016). Pendekatan dari penentuan k tersebut digunakan pada penelitian ini untuk menentukan nilai k yang digunakan, Jika dirumuskan, pendekatan aturan empiris tersebut berupa:

$k = \sqrt{n}$

n adalah jumlah dataset yang akan digunakan dalam melakukan klasifikasi, sehingga rumus aturan tersebut pada penelitian ini adalah :

$k = \sqrt{1000}$

$k = 31.6227766017$

Nilai k yang didapatkan dari rumus tersebut adalah 31.6227766017 sehingga dibulatkan menjadi 32. Nilai k yang digunakan dalam klasifikasi algoritma KNN penelitian ini adalah k=32.

4.6. Klasifikasi Menggunakan Naive Bayes

Proses lainnya setelah melakukan proses preprocessing dan seleksi fitur yang dibutuhkan adalah proses klasifikasi menggunakan algoritma Naive Bayes. Tahap ini akan menerapkan algoritma Naive Bayes yang telah dipaparkan pada sub-bab sebelumnya.

4.7. Uji dan Evaluasi Hasil

Dalam proses penelitian dari melakukan scraping hingga melakukan klasifikasi dengan menggunakan Algoritma KNN dan NB, peneliti menggunakan perangkat komputer yang dirakit dengan spesifikasi Processor Intel Core Intel® Core™ i5-12400F, RAM 32GB DDR4 3200MHz (Dual Channel kit), Solid State Drive 500 GB dan berjalan pada Sistem Operasi Microsoft Windows 10 Professional 22H2 64-bit.

Penelitian ini menggunakan data yang didapatkan dengan melakukan web scraping untuk mendapatkan sinopsis dari situs anime-manga database berbahasa inggris, terdapat 1000 data yang dikumpulkan dari scraping yang dibagi menjadi 250 sinopsis genre fantasy, 250 sinopsis genre mystery, 250 sinopsis genre romance, dan 250 sinopsis genre sports. Dari dataset yang digunakan, data dibagi menjadi proporsi 80% untuk data training dan 20% untuk data testing, proses pembagian data dilakukan secara acak dan setiap label mempunyai jumlah anggota yang sama saat dilakukan pembagian.

Data harus melalui tahap preprocessing, yaitu mengubah semua huruf menjadi huruf kecil (case folding), lalu menghapus teks yang tidak relevan (data cleansing), menghapus kata-kata umum (stopwords removal), menemukan kata-

kata ke bentuk dasarnya (stemming), dan pada tahap terakhir memisahkan semua kalimat menjadi kata-kata yang terpisah (tokenisasi). Masing-masing kombinasi akan dilakukan dan disimpan kedalam file .csv terpisah sesuai dengan penomoran kombinasi fitur yang digunakan. Kombinasi Stemming dan Stopwords removal dipilih karena kedua fitur tersebut dapat mengubah struktur bahasa dari sebuah teks.

Setelah tahap preprocessing, pembobotan TF-IDF dilakukan dengan menggunakan rumus (2), (3), dan (4), setelah bobot didapatkan, dilakukan seleksi fitur terlebih dahulu menggunakan Mutual Infomation untuk mengurangi fitur kata yang kurang relevan.

Klasifikasi dilakukan dengan menggunakan algoritma K-Nearest Neighbors (KNN). Nilai k yang digunakan dalam penelitian ini mengikuti pendekatan aturan umum yang ada yaitu akar dari jumlah data sampel yang menghasilkan k=32. Selain menggunakan KNN, digunakan pula algoritma Naive Bayes untuk melakukan klasifikasi dan membandingkan performa dari kedua algoritma klasifikasi tersebut.

Untuk mengevaluasi performa kedua algoritma tersebut, dilakukan perhitungan total data yang diklasifikasikan dengan benar dibagi oleh jumlah keseluruhan data uji. Informasi mengenai hasil klasifikasi ditampilkan dalam Tabel confusion matrix yang memberikan detail prediksi algoritma klasifikasi tersebut dengan perubahan kombinasi fitur dan penggunaan seleksi fitur.

Evaluasi dilaksanakan untuk menilai apakah penelitian telah mencapai tujuannya atau belum. Evaluasi penelitian ini melibatkan perhitungan akurasi, presisi, recall, dan F1-Score berdasarkan hasil klasifikasi yang terdokumentasikan

dalam confusion matrix. Berikut ini adalah hasil klasifikasi yang terdokumentasikan ke dalam confusion matriks

4.7.1. Klasifikasi K-Nearest Neighbors tanpa Mutual Information

Tabel-tabel dibawah ini adalah adalah confusion matrix untuk empat kombinasi text preprocessing yang berbeda dalam algoritma klasifikasi K-Nearest Neighbors (KNN) tanpa menggunakan Mutual Information. Setiap confusion matrix mewakili hasil dari prediksi terhadap kategori aktual untuk empat genre yang berbeda: Fantasy, Mystery, Romance, dan Sports.

Setiap confusion matriks menunjukkan jumlah prediksi yang tepat dan tidak tepat untuk setiap kategori aktual. Sebagai contoh, pada hasil di Tabel 4.9 (Kombinasi 1):

- Pada baris Fantasy, kolom Fantasy, angka 33 menunjukkan jumlah prediksi yang benar-benar adalah genre "Fantasy" yang sesuai dengan kategori aktualnya juga adalah "Fantasy".
- Pada baris Fantasy, kolom Mystery, angka 3 menunjukkan bahwa ada 3 prediksi yang sebenarnya adalah genre "Fantasy" tetapi diprediksi sebagai "Mystery".

Tabel 4.9. Confusion Matrix Kombinasi 1 Algoritma KNN tanpa Mutual Information

Kombinasi 1 – Tanpa Stopwords removal dan stemming Algoritma : K-Nearest Neighbors (k=32)				
Prediksi	Fantasy	Mystery	Romance	Sports
Aktual				

Tabel 4.9. Lanjutan Confusion Matrix Kombinasi 1 Algoritma KNN tanpa Mutual Information

Fantasy	33	3	6	8
Mystery	4	34	6	6
Romance	5	3	34	8
Sports	1	0	5	44

Tabel 4.10. Confusion Matrix Kombinasi 2 Algoritma KNN tanpa Mutual Information

Kombinasi 2 –Stopwords removal Algoritma : K-Nearest Neighbors (k=32)				
Prediksi	Fantasy	Mystery	Romance	Sports
Aktual				
Fantasy	36	4	6	4
Mystery	7	34	5	4
Romance	6	4	33	7
Sports	3	2	3	42

Tabel 4.11. Confusion Matrix Kombinasi 3 Algoritma KNN tanpa Mutual Information

Kombinasi 3 – Stemming Algoritma : K-Nearest Neighbors (k=32)				
Prediksi	Fantasy	Mystery	Romance	Sports
Aktual				
Fantasy	34	4	6	6

Tabel 4.11. Lanjutan Confusion Matrix Kombinasi 3 Algoritma KNN tanpa Mutual Information

Mystery	4	38	5	3
Romance	4	2	37	7
Sports	1	0	4	45

Tabel 4.12. Confusion Matrix Kombinasi 4 Algoritma KNN tanpa Mutual Information

Kombinasi 4 – Stopwords removal dan stemming Algoritma : K-Nearest Neighbors (k=32)				
Prediksi	Fantasy	Mystery	Romance	Sports
Aktual				
Fantasy	39	3	4	4
Mystery	7	37	4	2
Romance	4	4	36	6
Sports	2	1	3	44

Tabel 4.9. hingga Tabel 4.12. adalah Confusion Matrix yang membandingkan prediksi dengan kategori aktual dari beberapa jenis genre (Fantasy, Mystery, Romance, Sports). Setiap sel dalam tabel menyatakan jumlah prediksi tertentu yang cocok dengan kategori aktual tertentu. Jadi, tabel tersebut memberikan gambaran tentang seberapa baik prediksi genre untuk masing-masing kategori aktual.

Dari tabel-tabel yang diberikan, terlihat confusion matrix untuk empat kombinasi dari algoritma KNN tanpa menggunakan Mutual Information, dengan

variasi penghapusan stopwords, penggunaan stemming, atau keduanya. Confusion matrix ini memberikan informasi tentang seberapa baik model melakukan klasifikasi untuk setiap kategori. Berikut adalah beberapa temuan dari confusion matrix ini:

- Kombinasi 1 - Tanpa Stopwords Removal dan Stemming:

  Dalam confusion matrix ini, terlihat bahwa kelas "Sports" memiliki akurasi yang cukup baik, dengan angka prediksi yang tinggi dan kesalahan yang relatif rendah. Namun, kelas "Fantasy", "Romance", dan "Mystery" memiliki beberapa kebingungan, dengan beberapa sampel yang salah diklasifikasikan ke dalam kelas lain.

- Kombinasi 2 - Stopwords Removal:

  Dengan penghapusan stopwords, confusion matrix ini menunjukkan peningkatan dalam klasifikasi kelas "Romance" dan "Mystery". Namun, kelas "Fantasy" dan "Sports" masih memiliki klasifikasi yang lebih baik.

- Kombinasi 3 - Stemming:

  Dengan stemming, confusion matrix ini menunjukkan peningkatan dalam klasifikasi kelas "Mystery". Namun, kelas lainnya tampaknya memiliki performa yang hampir serupa dengan Kombinasi 1.

- Kombinasi 4 - Stopwords Removal dan Stemming:

  Kombinasi penghapusan stopwords dan stemming menunjukkan peningkatan dalam klasifikasi untuk beberapa kelas, terutama "Romance" dan "Mystery". Ini menunjukkan bahwa kedua metode pre-processing ini dapat memberikan kontribusi positif terhadap klasifikasi.

Dengan demikian, dari temuan ini, dapat disimpulkan bahwa penggunaan teknik pre-processing seperti penghapusan stopwords dan stemming (Kombinasi 4) dapat membantu meningkatkan kinerja model klasifikasi seperti K-Nearest Neighbors dalam mengklasifikasikan teks ke dalam kategori yang sesuai.

4.7.2. Klasifikasi Naive Bayes tanpa Mutual Information

Tabel-tabel dibawah ini adalah adalah confusion matrix untuk empat kombinasi text preprocessing yang berbeda dalam algoritma klasifikasi Naive Bayes tanpa menggunakan Mutual Information. Setiap confusion matrix mewakili hasil dari prediksi terhadap kategori aktual untuk empat genre yang berbeda: Fantasy, Mystery, Romance, dan Sports.

Tabel 4.13. Confusion Matrix Kombinasi 1 Algoritma Naive Bayes tanpa Mutual Information

Kombinasi 1 – Tanpa Stopwords removal dan stemming Algoritma : Naive Bayes				
Prediksi	Fantasy	Mystery	Romance	Sports
Aktual				
Fantasy	38	0	12	0
Mystery	10	27	13	0
Romance	3	1	44	2
Sports	1	0	4	45

Tabel 4.14. Confusion Matrix Kombinasi 2 Algoritma Naive Bayes tanpa Mutual Information

Kombinasi 2 –Stopwords removal Algoritma : Naive Bayes				
Prediksi	Fantasy	Mystery	Romance	Sports
Aktual				
Fantasy	44	2	3	1
Mystery	7	33	36	1
Romance	8	1	36	5
Sports	3	0	2	45

Tabel 4.15. Confusion Matrix Kombinasi 3 Algoritma Naive Bayes tanpa Mutual Information

Kombinasi 3 – Stemming Algoritma : Naive Bayes				
Prediksi	Fantasy	Mystery	Romance	Sports
Aktual				
Fantasy	37	2	11	0
Mystery	9	33	8	0
Romance	2	1	45	2
Sports	1	0	4	45

Tabel 4.16. Confusion Matrix Kombinasi 4 Algoritma Naive Bayes tanpa Mutual Information

Kombinasi 4 – Stopwords removal dan stemming Algoritma : Naive Bayes				
Prediksi	Fantasy	Mystery	Romance	Sports
Aktual				
Fantasy	37	3	8	2
Mystery	10	38	1	1
Romance	2	2	44	2
Sports	1	1	3	45

Tabel 4.13. hingga Tabel 4.16. adalah Confusion Matrix yang membandingkan prediksi dengan kategori aktual dari beberapa jenis genre (Fantasy, Mystery, Romance, Sports). Setiap sel dalam tabel menyatakan jumlah prediksi tertentu yang cocok dengan kategori aktual tertentu. Jadi, tabel tersebut memberikan gambaran tentang seberapa baik prediksi genre untuk masing-masing kategori aktual.

Dari tabel-tabel yang ditampilkan, masing-masing confusion matrix menunjukkan hasil dari empat kombinasi algoritma Naive Bayes tanpa menggunakan Mutual Information, dengan variasi penghapusan stopwords, penggunaan stemming, atau keduanya. Berikut adalah beberapa temuan dari confusion matrix ini:

- Kombinasi 1 - Tanpa Stopwords Removal dan Stemming:

Terdapat kesulitan dalam mengklasifikasikan kelas "Fantasy" dan "Mystery", dengan beberapa sampel yang salah diklasifikasikan ke dalam kelas "Romance". Hal ini terlihat dari jumlah false positives yang cukup tinggi untuk kelas "Romance".

- Kombinasi 2 - Stopwords Removal:

  Penghapusan stopwords tampaknya sedikit meningkatkan klasifikasi kelas "Mystery", dengan jumlah false negatives yang sedikit berkurang dibandingkan dengan Kombinasi 1.

- Kombinasi 3 - Stemming:

  Stemming tidak tampak memberikan dampak signifikan pada kinerja klasifikasi dibandingkan dengan Kombinasi 1 dan Kombinasi 2. False Negatives untuk klasifikasi kelas "Mystery" masih cukup tinggi

- Kombinasi 4 - Stopwords Removal dan Stemming:

  Kombinasi penghapusan stopwords dan stemming memberikan hasil signifikan untuk kelas "Mystery", Namun, masih terlihat beberapa kesulitan dalam mengklasifikasikan kelas "Mystery" karena beberapa sampel "Mystery" yang salah diklasifikasikan ke dalam kelas "Fantasy".

Dari temuan ini, dapat disimpulkan bahwa penggunaan teknik pre-processing seperti penghapusan stopwords dan stemming memiliki efek pada kinerja algoritma Naive Bayes dalam hal ini. Namun, kelas "Mystery" tampaknya menjadi tantangan dalam klasifikasi dan mungkin memerlukan pendekatan tambahan untuk meningkatkan akurasi klasifikasi.

4.7.3. Nilai Performa Tanpa Mutual Information

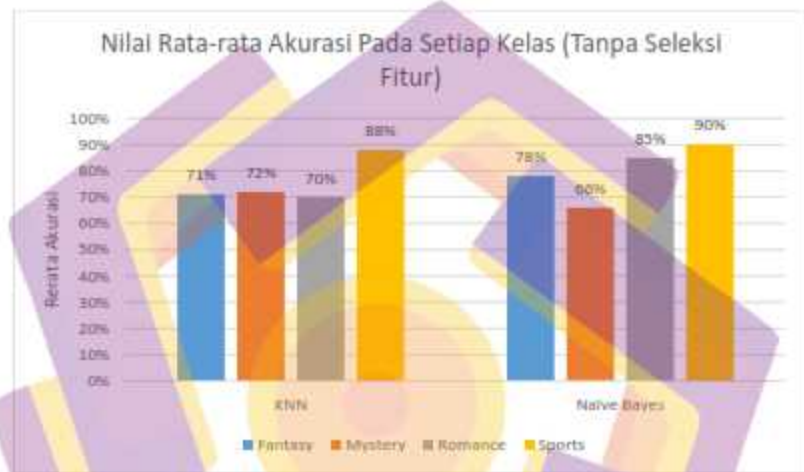
Berdasarkan hasil confusion matrix pada subbab sebelumnya, Tabel ini menunjukkan hasil akurasi dari algoritma klasifikasi teks K-Nearest Neighbors (KNN) dengan =32 dan Naive Bayes pada empat kelas teks yang berbeda, yaitu Fantasy, Mystery, Romance, dan Sports. Kedua algoritma diuji pada empat kombinasi pengolahan data teks: tanpa stemming dan stopwords removal, dengan stopwords removal, dengan stemming, serta dengan kombinasi stemming dan stopwords removal.

Tabel 4.17. Tabel Pengukuran Nilai Akurasi Pada Setiap Genre Tanpa Mutual Information

Algoritma	Kelas	Akurasi Pada Kombinasi 1	Akurasi Pada Kombinasi 2	Akurasi Pada Kombinasi 3	Akurasi Pada Kombinasi 4	Rata-Rata
KNN (k=32)	Fantasy	66%	72%	68%	78%	**71%**
	Mystery	68%	68%	76%	74%	**72%**
	Romance	68%	66%	74%	72.0%	**70%**
	Sports	88%	84%	90%	88%	**88%**
Naive Bayes	Fantasy	76%	88%	74%	74%	**78%**
	Mystery	54%	66%	66%	76%	**66%**

Tabel 4.17. Lanjutan Tabel Pengukuran Nilai Akurasi Pada Setiap Genre Tanpa Mutual Information

	Romance	88%	72%	90%	88%	**85%**
	Sports	90%	90%	90%	90%	**90%**



Gambar 4.4. Grafik Nilai Rata-Rata Akurasi pada Setiap Genre (Tanpa Seleksi Fitur)

Dari Tabel 4.17 dan Gambar 4.4 dapat disimpulkan bahwa genre Sports memiliki akurasi paling tinggi dari seluruh pengujian, baik untuk algoritma K-Nearest Neighbors (KNN) maupun Naive Bayes. Genre ini mencapai akurasi tertinggi dan paling konsisten, yaitu 84%-90% di semua kombinasi preprocessing data untuk KNN, dan 90% di semua kombinasi untuk Naive Bayes.

Alasan mengapa genre Sports memiliki akurasi yang paling tinggi kemungkinan besar adalah karena teks dalam kelas ini lebih konsisten dalam pola dan terminologi yang digunakan. Bahasa dan kata-kata yang sering muncul dalam

sinopsis bertema olahraga mungkin lebih mudah untuk diklasifikasikan oleh algoritma-algoritma tersebut. Faktor lain bisa jadi adalah kurangnya ambiguitas dalam kata-kata yang memeliki keterkaitan dengan sinopsis olahraga dibandingkan dengan genre lain seperti Mystery, di mana variasi bahasa dan gaya penulisan lebih besar. Dengan kata lain, karakteristik linguistik dan struktur konten dalam teks olahraga mungkin lebih homogen, sehingga lebih mudah untuk dikenali dan diklasifikasikan dengan tepat oleh algoritma machine learning tersebut.

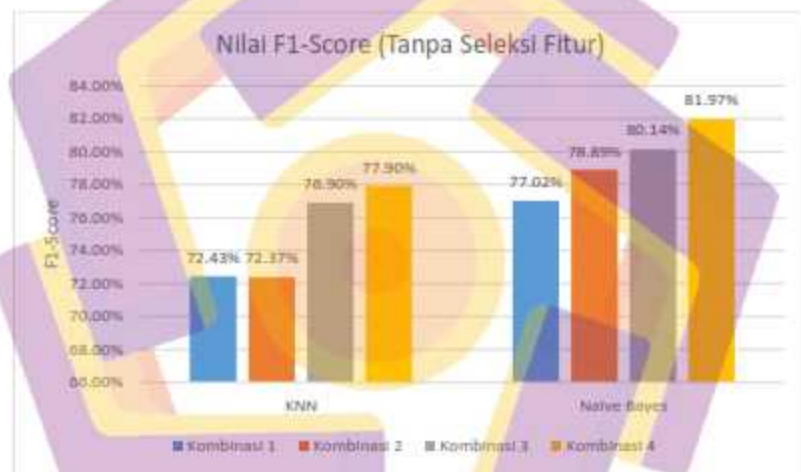
Tabel 4.18. dibawah ini merupakan hasil evaluasi performa dari dua algoritma klasifikasi, yaitu KNN (K-Nearest Neighbors) dan Naive Bayes, menggunakan empat kombinasi yang berbeda dalam pengolahan teks tanpa seleksi fitur, Nilai performa dari setiap algoritma dievaluasi menggunakan metrik seperti akurasi, presisi, recall, dan F1-score. Metrik-metrik ini memberikan informasi tambahan tentang seberapa baik model klasifikasi bekerja.

Tabel 4.18. Tabel Pengukuran Nilai Performa Tanpa Mutual Information

Algoritma	Matrix	Kombinasi 1 (Tanpa Stemming dan Stopwords Removal	Kombinasi 2 (Stopwords Removal)	Kombinasi 3 (Stemming)	Kombinasi 4 (Stemming + Stopwords Removal)
KNN (k=32)	Accuracy	72.5%	72.5%	77.0%	**78.0%**
	Precision	73.769%	72.6%	77.589%	**78.097%**
	Recall	72.5%	72.5%	77%	**78.0%**

Tabel 4.18. Lanjutan Tabel Pengukuran Nilai Performa Tanpa Mutual Information

	F1-Score	72.428%	72.369%	76.9%	**77.902%**
Naive Bayes	Accuracy	77.0%	79.0%	80.0%	**82.0%**
	Precision	81.381%	80.293%	82.274%	**82.234%**
	Recall	77.0%	79.0%	80.0%	**82.0%**
	F1-Score	77.017%	78.89%	80.136%	**81.967%**



Gambar 4.5. Grafik Nilai F1-Score Pengujian tanpa Seleksi Fitur

Dalam Gambar 4.5. Untuk KNN, terlihat bahwa penggunaan Stemming saja (Kombinasi 3) atau kombinasi Stemming dengan Stopwords Removal (Kombinasi 4) meningkatkan akurasi dan terjadi peningkatan yang konsisten dalam precision, recall, dan F1-Score dibandingkan dengan tidak menggunakan keduanya, Ini menunjukkan bahwa dalam kasus ini, penggunaan Stemming dapat membantu meningkatkan kinerja model untuk mengidentifikasi dan memulihkan informasi dengan lebih baik.

Dalam Gambar 4.5 Pada kasus Naive Bayes, terlihat bahwa penggunaan kombinasi Stemming dengan Stopwords Removal (Kombinasi 4) memberikan peningkatan yang signifikan dalam semua metrik evaluasi dibandingkan dengan kombinasi lainnya. Sedikit Berbeda dengan KNN, kombinasi 2 menghasilkan nilai F1-Score yang menurun tipis daripada kombinasi 1. Ini menunjukkan bahwa penggunaan Stemming dengan penghapusan stopwords dapat meningkatkan performa Naive Bayes dalam kasus ini.

Terlihat dalam tabel 4.18 bahwa performa terbaik untuk klasifikasi yang tidak menggunakan seleksi fitur, ditunjukkan oleh Algoritma Naive Bayes menggunakan kombinasi 4 dengan hasil akurasi 82.0%, precision 82.234%, recall 82.0%, F1-Score 81.967%, hasil ini lebih tinggi dari hasil terbaik algoritma KNN menggunakan kombinasi 4 yang hanya mendapatkan nilai akurasi 78.0%, precision 78.087% , recall 78.0% , F1-score 77.902%. Dari hasil ini dapat dikatakan bahwa Naive Bayes mengungguli KNN dalam mengklasifikasikan genre anime berdasarkan sinopsis dan Kombinasi Fitur 4 memiliki hasil yang paling baik diantara kombinasi fitur yang lainnya. Kesalahan atau kegagalan dalam klasifikasi terjadi karena kemiripan antara kata-kata yang membentuk suatu kelas dengan kata-kata dalam kelas lain serta kata-kata dari setiap data yang digunakan.

Dengan demikian, tabel ini memberikan wawasan tentang bagaimana penggunaan teknik-teknik pemrosesan teks tertentu (seperti Stemming dan penghapusan stopwords) dapat mempengaruhi performa dari algoritma-algoritma klasifikasi menggunakan KNN dan Naive Bayes.

4.7.4. Klasifikasi K-Nearest Neighbors dengan Mutual Information

Eksperimen selanjutnya adalah menggunakan Seleksi fitur Mutual Information untuk menghilangkan fitur-fitur yang tidak penting untuk proses klasifikasi. Tabel-tabel dibawah ini adalah adalah confusion matrix untuk empat kombinasi text preprocessing yang berbeda dalam algoritma klasifikasi KNN dengan menggunakan Mutual Information. Setiap confusion matrix mewakili hasil dari prediksi terhadap kategori aktual untuk empat genre yang berbeda: Fantasy, Mystery, Romance, dan Sports.

Tabel 4.19. Confusion Matrix Kombinasi 1 Algoritma KNN dengan Mutual Information

Kombinasi 1 – Tanpa Stopwords removal dan stemming Algoritma : K-Nearest Neighbors (k=32)				
Prediksi	Fantasy	Mystery	Romance	Sports
Aktual				
Fantasy	42	4	4	0
Mystery	16	30	4	0
Romance	13	3	33	1
Sports	8	1	9	32

Tabel 4.20. Confusion Matrix Kombinasi 2 Algoritma KNN dengan Mutual Information

Kombinasi 2 –Stopwords removal Algoritma : K-Nearest Neighbors (k=32)				
Prediksi	Fantasy	Mystery	Romance	Sports

Tabel 4.20. Lanjutan Confusion Matrix Kombinasi 2 Algoritma KNN dengan Mutual Information

Aktual				
Fantasy	32	9	8	1
Mystery	5	41	4	0
Romance	3	14	31	2
Sports	2	7	6	35

Tabel 4.21. Confusion Matrix Kombinasi 3 Algoritma KNN dengan Mutual Information

Kombinasi 3 – Stemming Algoritma : K-Nearest Neighbors (k=32)				
Prediksi	Fantasy	Mystery	Romance	Sports
Aktual				
Fantasy	47	0	3	0
Mystery	17	33	0	0
Romance	19	4	26	1
Sports	5	4	5	36

Tabel 4.22. Confusion Matrix Kombinasi 4 Algoritma KNN dengan Mutual Information

Kombinasi 4 – Stopwords removal dan stemming Algoritma : K-Nearest Neighbors (k=32)				
Prediksi	Fantasy	Mystery	Romance	Sports
Aktual				

Tabel 4.22. Lanjutan Confusion Matrix Kombinasi 4 Algoritma KNN dengan Mutual Information

Fantasy	44	4	1	1
Mystery	8	34	7	1
Romance	11	7	31	1
Sports	5	5	4	36

Tabel 4.19. hingga Tabel 4.22. adalah masing-masing confusion matrix menunjukkan hasil dari empat kombinasi preprocessing untuk algoritma K-Nearest Neighbors (KNN) dengan penggunaan Mutual Information, dengan variasi penghapusan stopwords, penggunaan stemming, atau keduanya. Berikut adalah beberapa temuan dari confusion matrix ini:

- Kombinasi 1 - Tanpa Stopwords Removal dan Stemming

  Kombinasi ini menunjukkan kinerja kurang baik dalam mengklasifikasikan keempat kelas, dengan beberapa kesalahan terutama terjadi pada kelas "Mystery" dan "Romance" yang banyak sampel diklasifikasikan ke dalam kelas "Fantasy"

- Kombinasi 2 - Stopwords Removal

  Penghapusan stopwords tampaknya sedikit meningkatkan kinerja klasifikasi, terutama untuk kelas "Mystery" dan "Sports". Namun, kelas "Fantasy" dan "Romance" masih memiliki beberapa kesalahan.

- Kombinasi 3 - Stemming

  Stemming memberikan dampak signifikan pada kinerja klasifikasi, tetapi kelas "Fantasi" memiliki false positive yang tinggi.

- Kombinasi 4 - Stopwords Removal dan Stemming:

  Kombinasi penghapusan stopwords dan stemming memberikan hasil yang sedikit lebih baik dari kombinasi lainnya, tetapi tetap saja terjadi beberapa kesalahan terutama terjadi pada kelas "Mystery" dan "Romance" yang banyak sampel diklasifikasikan ke dalam kelas "Fantasy"

Namun, penggunaan teknik pre-processing seperti penghapusan stopwords atau stemming tidak memberikan perubahan yang signifikan dalam kinerja algoritma yang datasetnya dilakukan seleksi fitur terlebih dahulu. Terdapat beberapa kesalahan yang terjadi terutama pada kelas-kelas yang memiliki overlap dalam ciri-ciri teks mereka, seperti "Mystery" dan "Romance".

4.7.5. Klasifikasi Naive Bayes dengan Mutual Information

Eksperimen selanjutnya adalah menggunakan Seleksi fitur Mutual Information untuk menghilangkan fitur-fitur yang tidak penting untuk proses klasifikasi. Tabel-tabel dibawah ini adalah adalah confusion matrix untuk empat kombinasi text preprocessing yang berbeda dalam algoritma klasifikasi Naive Bayes dengan menggunakan Mutual Information. Setiap confusion matrix mewakili hasil dari prediksi terhadap kategori aktual untuk empat genre yang berbeda: Fantasy, Mystery, Romance, dan Sports.

Tabel 4.23. Confusion Matrix Kombinasi 1 Algoritma Naive Bayes dengan Mutual Information

Kombinasi 1 – Tanpa Stopwords removal dan stemming Algoritma : Naive Bayes				
Prediksi	Fantasy	Mystery	Romance	Sports

Tabel 4.23. Lanjutan Confusion Matrix Kombinasi 1 Algoritma Naive Bayes dengan Mutual Information

Aktual				
Fantasy	40	1	8	1
Mystery	14	25	10	1
Romance	5	0	44	1
Sports	2	2	4	42

Tabel 4.24. Confusion Matrix Kombinasi 2 Algoritma Naive Bayes dengan Mutual Information

Kombinasi 2 –Stopwords removal Algoritma : Naive Bayes				
Prediksi	Fantasy	Mystery	Romance	Sports
Aktual				
Fantasy	41	1	6	2
Mystery	12	30	8	0
Romance	7	3	38	2
Sports	3	2	4	41

Tabel 4.25. Confusion Matrix Kombinasi 3 Algoritma Naive Bayes dengan Mutual Information

Kombinasi 3 – Stemming Algoritma : Naive Bayes				
Prediksi	Fantasy	Mystery	Romance	Sports
Aktual				

Tabel 4.25. Lanjutan Confusion Matrix Kombinasi 3 Algoritma Naive Bayes dengan Mutual Information

Fantasy	38	1	9	2
Mystery	11	32	7	0
Romance	3	1	44	2
Sports	2	2	4	42

Tabel 4.26. Confusion Matrix Kombinasi 4 Algoritma Naive Bayes dengan Mutual Information

Kombinasi 4 – Stopwords removal dan stemming Algoritma : Naive Bayes				
Prediksi	Fantasy	Mystery	Romance	Sports
Aktual				
Fantasy	38	5	5	2
Mystery	7	38	3	2
Romance	1	4	42	3
Sports	1	3	4	42

Tabel 4.23. hingga Tabel 4.26. adalah masing-masing confusion matrix menunjukkan hasil dari empat kombinasi preprocessing untuk algoritma Naive Bayes dengan penggunaan Mutual Information, dengan variasi penghapusan stopwords, penggunaan stemming, atau keduanya. Berikut adalah beberapa temuan dari confusion matrix ini:

- Kombinasi 1 - Tanpa Stopwords Removal dan Stemming:

Kombinasi ini menunjukkan kinerja yang kurang baik dalam mengklasifikasikan keempat kelas, terdapat banyak kesalahan yang terutama terjadi pada kelas "Mystery".

- Kombinasi 2 - Stopwords Removal:

  Penghapusan stopwords tampaknya tidak meningkatkan kinerja klasifikasi, kelas "Mystery" masih memiliki beberapa kesalahan dan diikuti oleh kelas "Romance" yang semakin banyak menghasilkan false negatives

- Kombinasi 3 - Stemming:

  Stemming tidak memberikan dampak signifikan pada kinerja klasifikasi, walaupun seperti itu tampaknya kelas "Mystery" dan kelas "Romance" memiliki false negatives yang lebih sedikit daripada kombinasi sebelumnya.

- Kombinasi 4 - Stopwords Removal dan Stemming:

  Kombinasi penghapusan stopwords dan stemming memberikan hasil yang cukup baik. Namun, masih terdapat beberapa kesalahan terutama pada kelas "Mystery" dan "Fantasy". Kelas "Romance" mendapatkan hasil yang signifikan pada kombinasi ini

Dari temuan ini, dapat disimpulkan bahwa penggunaan Mutual Information dalam algoritma Naive Bayes memberikan kinerja yang baik dalam mengklasifikasikan teks ke dalam kategori yang sesuai. Penggunaan teknik pre-processing seperti penghapusan stopwords atau stemming memiliki pengaruh yang bervariasi tergantung pada dataset dan kelas-kelas yang ada. Terdapat beberapa

kesalahan yang terjadi terutama pada kelas yang memiliki overlap dalam ciri-ciri teks, seperti kelas "Mystery".

4.7.6. Nilai Performa Dengan Mutual Information

Berdasarkan hasil confusion matrix pada subbab sebelumnya, Tabel ini menunjukkan hasil akurasi dari algoritma klasifikasi teks K-Nearest Neighbors (KNN) dengan =32 dan Naive Bayes dengan mutual information pada empat kelas teks yang berbeda, yaitu Fantasy, Mystery, Romance, dan Sports. Kedua algoritma diuji pada empat kombinasi pengolahan data teks: tanpa stemming dan stopwords removal, dengan stopwords removal, dengan stemming, serta dengan kombinasi stemming dan stopwords removal.

Tabel 4.27. Tabel Pengukuran Nilai Akurasi Pada Setiap Genre Dengan Mutual Information

Algoritma	Kelas	Akurasi Pada Kombinasi 1 (Tanpa Stemming dan Stopwords Removal	Akurasi Pada Kombinasi 2 (Stopwords Removal)	Akurasi Pada Kombinasi 3 (Stemming)	Akurasi Pada Kombinasi 4 (Stemming + Stopwords Removal)	Rata-Rata
KNN (K=32)	Fantasy	84%	64%	94%	88%	**83%**
	Mystery	60%	82%	66%	68%	**69%**

Tabel 4.27. Lanjutan Tabel Pengukuran Nilai Akurasi Pada Setiap Genre Dengan Mutual Information

+ Seleksi Fitur	Romance	66%	62%	52%	62%	**61%**
	Sports	64%	70%	72%	72%	**70%**
Naive Bayes + Seleksi Fitur	Fantasy	80%	82%	76%	76%	**79%**
	Mystery	50%	60%	64%	76%	**63%**
	Romance	88%	76%	88%	84%	**84%**
	Sports	84%	82%	84%	84%	**84%**

Gambar 4.6. Grafik Nilai Rata-Rata Akurasi pada Setiap Genre (Dengan Seleksi Fitur)

Dari Tabel 4.27 dan Gambar 4.6 dapat dikatakan bahwa Kelas Fantasy menjadi kelas yang memiliki akurasi paling tinggi dari pengujian menggunakan algoritma KNN dan seleksi fitur dengan rata-rata akurasi sebesar 83%.

Sementara itu, Kelas Sports dan Romance memiliki akurasi paling tinggi dari seluruh pengujian menggunakan algoritma Naive Bayes dan seleksi fitur dengan rata-rata akurasi 84%. algoritma Naive Bayes bekerja sangat baik pada kategori Sports dan Romance karena adanya kosakata yang spesifik dan konsisten, distribusi kata yang jelas, dan efektivitas teknik preprocessing yang membantu memperkuat fitur-fitur relevan. Naive Bayes memanfaatkan probabilitas tinggi dari kata-kata yang unik untuk setiap kategori ini, membuatnya lebih mudah untuk mengklasifikasikan teks dengan benar dan mencapai akurasi yang tinggi.

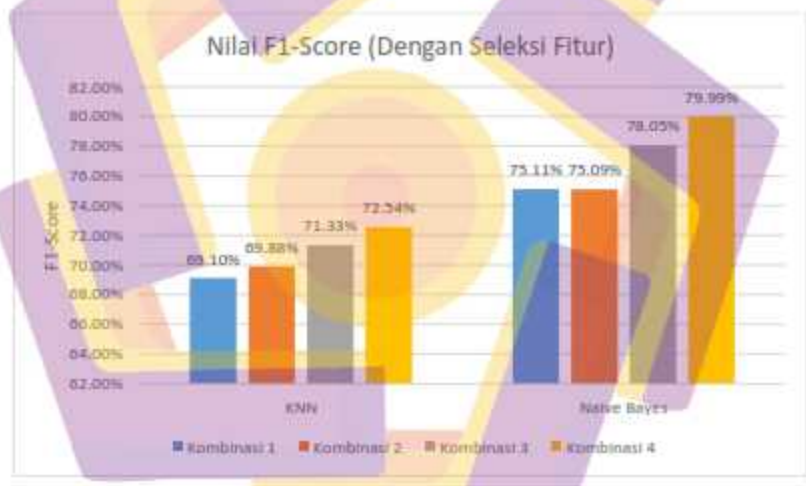
Tabel 4.28 ini merupakan hasil evaluasi performa dari dua algoritma klasifikasi, yaitu KNN (K-Nearest Neighbors) dan Naive Bayes, menggunakan empat kombinasi yang berbeda dalam pengolahan teks, Nilai performa dari setiap algoritma dievaluasi menggunakan metrik seperti akurasi, presisi, recall, dan F1-score. Metrik-metrik ini memberikan informasi tambahan tentang seberapa baik model klasifikasi bekerja.

Tabel 4.28. Tabel Pengukuran Nilai Performa Dengan Mutual Information

Algoritma	Matrix	Kombinasi 1 (Tanpa Stemming dan Stopwords Removal	Kombinasi 2 (Stopwords Removal)	Kombinasi 3 (Stemming)	Kombinasi 4 (Stemming + Stopwords Removal)
KNN (k=32)	Accuracy	68.5%	69.5%	71.0%	**72.5%**
	Precision	73.77%	72.327%	76.916%	**74.277%**

Tabel 4.28. Lanjutan Tabel Pengukuran Nilai Performa Dengan Mutual Information

+ Seleksi Fitur	Recall	68.5%	69.5%	71.0%	**72.5%**
	F1-Score	69.102%	69.876%	71.327%	**72.535%**
Naive Bayes + Seleksi Fitur	Accuracy	75.5%	75.0%	78.0%	**80.0%**
	Precision	78.715%	76.845	79.828%	**80.085%**
	Recall	75.5%	75.0%	78.0%	**80.0%**
	F1-Score	75.114%	75.087%	78.047%	**79.992%**



Gambar 4.7. Grafik Nilai F1-Score Pengujian dengan Seleksi Fitur

Dalam Gambar 4.7. Untuk KNN, peningkatan dalam akurasi ketika menggunakan berbagai kombinasi preprocessing teks. Akurasi meningkat dari 68.5% (Kombinasi 1) menjadi 72.5% (Kombinasi 4), menunjukkan bahwa penggunaan kombinasi stemming dan stopwords removal memberikan peningkatan yang signifikan. Secara umum, terdapat peningkatan dalam precision, recall, dan

F1-Score ketika menggunakan kombinasi preprocessing teks yang lebih lengkap (Stemming, Stopwords Removal, atau kombinasi keduanya). Hal ini menunjukkan bahwa preprocessing teks yang lebih baik dapat membantu model KNN untuk melakukan prediksi yang lebih baik.

Mirip dengan KNN, Naive Bayes juga menunjukkan peningkatan dalam akurasi dengan penggunaan preprocessing teks yang lebih lengkap. Akurasi meningkat dari 75.5% (Kombinasi 1) menjadi 80.0% (Kombinasi 4), menunjukkan bahwa penggunaan stemming dan stopwords removal juga memberikan peningkatan yang signifikan di sini. precision, recall, dan F1-Score Naive Bayes juga meningkat dengan penggunaan preprocessing teks yang lebih baik.

Terlihat dalam tabel 4.28. bahwa performa terbaik untuk klasifikasi yang menggunakan seleksi fitur terlebih dahulu, ditunjukkan oleh Algoritma Naive Bayes menggunakan kombinasi 4 dengan hasil akurasi 80.0%, precision 80.085%, recall 80.0%, F1-Score 79.992%, hasil ini lebih tinggi dari hasil terbaik algoritma KNN menggunakan kombinasi 4 yang hanya mendapatkan nilai akurasi 72.5% , precision 74.277%, recall 72.5%, dan F1-score 72.535%. Dari hasil ini dapat dikatakan bahwa Naive Bayes dengan menggunakan seleksi fitur mutual information mengungguli KNN dengan menggunakan seleksi fitur mutual information dalam mengklasifikasikan genre anime berdasarkan sinopsis dan Kombinasi Fitur 4 memiliki hasil yang paling baik diantara kombinasi fitur yang lainnya. Kesalahan atau kegagalan dalam klasifikasi terjadi karena kemiripan antara kata-kata yang membentuk suatu kelas dengan kata-kata dalam kelas lain serta kata-kata dari setiap data yang digunakan.

4.7.6. Komparasi Hasil

Tabel ini menyajikan hasil tentang bagaimana dua algoritma klasifikasi pada penelitian ini yaitu KNN (dengan k=32) dan Naive Bayes bekerja dengan dan tanpa seleksi fitur pada empat kelas(genre) data yaitu Fantasy, Mystery, Romance, dan Sports. Dari tabel, terlihat bahwa seleksi fitur memiliki pengaruh yang berbeda tergantung pada kategori data dan algoritma yang digunakan.

Tabel 4.29. Tabel Pengukuran Rata-Rata Nilai Akurasi Pada Setiap Genre Untuk Seluruh Skenario Pengujian

Algoritma	Kelas	Akurasi Pada Kombinasi 1	Akurasi Pada Kombinasi 2	Akurasi Pada Kombinasi 3	Akurasi Pada Kombinasi 4	**Rata-Rata**
KNN (k=32)	Fantasy	66%	72%	68%	78%	71%
	Mystery	68%	68%	76%	74%	72%
	Romance	68%	66%	74%	72.0%	70%
	Sports	88%	84%	90%	88%	**88%**
Naive Bayes	Fantasy	76%	88%	74%	74%	78%
	Mystery	54%	66%	66%	76%	66%
	Romance	88%	72%	90%	88%	85%
	Sports	90%	90%	90%	90%	**90%**
KNN (k=32)	Fantasy	84%	64%	94%	88%	**83%**
	Mystery	60%	82%	66%	68%	69%

Tabel 4.29. Lanjutan Tabel Pengukuran Rata-Rata Nilai Akurasi Pada Setiap Genre Untuk Seluruh Skenario Pengujian

+ Seleksi Fitur	Romance	66%	62%	52%	62%	61%
	Sports	64%	70%	72%	72%	70%
Naive Bayes + Seleksi Fitur	Fantasy	80%	82%	76%	76%	79%
	Mystery	50%	60%	64%	76%	63%
	Romance	88%	76%	88%	84%	**84%**
	Sports	84%	82%	84%	84%	**84%**

Gambar 4.8. Grafik Nilai Rata-Rata Akurasi Pada Setiap Genre Untuk Seluruh Skenario Pengujian

Berdasarkan tabel 4.29 dan gambar 4.8 yang disajikan, genre Sports menghasilkan rata-rata akurasi paling tinggi pada mayoritas skenario pengujian, yaitu 88% untuk KNN (k=32) dan 90% untuk Naive Bayes, keduanya tanpa seleksi

fitur. Jika menggunakan seleksi fitur, kelas Sports mengalami penurunan akurasi yaitu 70% untuk KNN dan 84% untuk Naive Bayes.

Alasan mengapa genre Sports memiliki akurasi yang paling tinggi kemungkinan besar adalah karena teks dalam kelas ini lebih konsisten dalam pola dan terminologi yang digunakan. Istilah dan konteks yang sering digunakan dalam olahraga biasanya sangat spesifik dan berbeda dari genre lain, sehingga algoritma dapat mengenali pola dengan lebih mudah dan tidak sering muncul dalam kelas lain. Faktor lain bisa jadi adalah kurangnya ambiguitas dalam genre sports dibandingkan dengan genre lain seperti Mystery, di mana variasi bahasa dan gaya penulisan lebih besar. Dengan kata lain, karakteristik linguistik dan struktur konten dalam genre Sports mungkin lebih homogen, sehingga lebih mudah untuk dikenali dan diklasifikasikan dengan tepat oleh algoritma machine learning yang digunakan dalam penelitian ini. Walaupun seperti itu, penurunan akurasi pada saat menggunakan seleksi fitur terjadi karena fitur yang dipilih mungkin tidak selalu mencakup seluruh aspek penting dari data. Terkadang, fitur yang tampak tidak signifikan dalam perhitungan mutual information masih memiliki kontribusi yang penting ketika dikombinasikan dengan fitur lain. Penghapusan fitur-fitur ini bisa mengurangi kemampuan algoritma pengujian untuk menangkap informasi yang relevan.

Sementara itu, genre Mystery memiliki rata-rata akurasi paling rendah dari seluruh pengujian. Hal ini kemungkinan disebabkan oleh beberapa faktor yang membuat data dalam genre Mystery lebih sulit untuk diklasifikasikan dengan akurasi tinggi. Salah satu alasan utama mungkin adalah variasi dan kompleksitas

dalam sinopsis Mystery yang lebih tinggi dibandingkan dengan kelas lain. Sinopsis Mystery bisa mencakup berbagai subgenre dan gaya yang berbeda, sehingga menciptakan lebih banyak ambiguitas dan kurangnya pola yang jelas untuk dikenali oleh algoritma. Selain itu, fitur-fitur dalam sinopsis Mystery mungkin tidak sejelas atau sepadat fitur dalam genre lain seperti Sports. Misalnya, elemen-elemen cerita dalam genre Mystery bisa lebih bervariasi dan kompleks, dengan petunjuk, plot twist, dan karakterisasi yang lebih beragam. Hal ini mengakibatkan genre ini lebih sulit bagi algoritma untuk menemukan dan memanfaatkan pola yang konsisten untuk klasifikasi.

Selanjutnya adalah hasil keseluruhan pengujian pada algoritma dan skenario pengujian yang diusulkan pada dataset peneliti. Tabel 4.30. merupakan hasil dari seluruh pengujian yang dilakukan dalam penelitian ini dengan skenario pengujian yang berbeda. Perbedaan Simulasi penelitian mencakup perbedaan Algoritma (KNN dan NB), Penggunaan Seleksi fitur, dan perbedaan kombinasi fitur preprocessing. Tabel ini merupakan hasil dari seluruh simulasi pengujian yang dilakukan oleh peneliti.

Tabel 4.30. Tabel Hasil keseluruhan pengujian dan Pengukuran Nilai Performa

Algoritma	Matrix	Kombinasi 1 (Tanpa Stemming dan Stopwords Removal	Kombinasi 2 (Stopwords Removal)	Kombinasi 3 (Stemming)	Kombinasi 4 (Stemming + Stopwords Removal)

Tabel 4.30. Lanjutan Tabel Hasil keseluruhan pengujian dan Pengukuran Nilai Performa

KNN (k=32)	Accuracy	72.5%	72.5%	77.0%	**78.0%**
	Precision	73.769%	72.6%	77.589%	**78.097%**
	Recall	72.5%	72.5%	77%	**78.0%**
	F1-Score	72.428%	72.369%	76.9%	**77.902%**
Naive Bayes	Accuracy	77.0%	79.0%	80.0%	**82.0%**
	Precision	81.381%	80.293%	82.274%	**82.234%**
	Recall	77.0%	79.0%	80.0%	**82.0%**
	F1-Score	77.017%	78.89%	80.136%	**81.967%**
KNN (k=32) + Seleksi Fitur	Accuracy	68.5%	69.5%	71.0%	**72.5%**
	Precision	73.77%	72.327%	76.916%	**74.277%**
	Recall	68.5%	69.5%	71.0%	**72.5%**
	F1-Score	69.102%	69.876%	71.327%	**72.535%**
Naive Bayes + Seleksi Fitur	Accuracy	75.5%	75.0%	78.0%	**80.0%**
	Precision	78.715%	76.845	79.828%	**80.085%**
	Recall	75.5%	75.0%	78.0%	**80.0%**
	F1-Score	75.114%	75.087%	78.047%	**79.992%**

Gambar 4.9. Grafik Nilai F1-Score Seluruh Pengujian

Terlihat pada gambar 4.9., Pada algoritma KNN (k=32), kombinasi 3 (Stemming) dan kombinasi 4 (Stemming + Stopwords Removal) menunjukkan peningkatan dalam semua metrik evaluasi (Accuracy, Precision, Recall, dan F1-Score) dibandingkan dengan kombinasi 1 (Tanpa Stemming dan Stopwords Removal) dan kombinasi 2 (Stopwords Removal). Sedangkan pada algoritma Naive Bayes, kombinasi kombinasi 2 (Stopwords Removal), 3 (Stemming) dan kombinasi 4 (Stemming + Stopwords Removal) juga menunjukkan peningkatan dalam semua metrik evaluasi dibandingkan dengan kombinasi 1 (Tanpa Stemming dan Stopwords Removal).

Penggunaan Stopwords Removal tidak memberikan dampak yang signifikan, terlihat hanya terjadi perubahan yang sangat kecil dalam hasil klasifikasi, walaupun begitu penurunan performa ketika menggunakan stopwords Removal terjadi karena banyaknya informasi yang hilang pada dataset setelah proses stopword removal. Proses stemming dapat membantu meningkatkan hasil

pada kedua algoritma tersebut dengan cara mengurangi variasi kata sehingga minim noise dan outliers, Namun efektivitasnya tergantung pada seberapa baik teknik ini mampu mempertahankan makna asli kata. Penggunaan proses Stemming yang dipadukan dengan Stopwords Removal menghasilkan nilai pengujian yang paling baik setidaknya untuk penelitian kasus ini. Hal ini terjadi karena beberapa kata setelah dilakukan Stemming akan berubah menjadi stopwords dan kemudian proses stopwords removal akan menghilangkan stopwords yang tidak diperlukan sehingga data menjadi lebih bersih dari noise atau outliers sehingga efektivitas kedua algoritma yang digunakan menjadi meningkat. Eksperimen ini menunjukkan bahwa fitur preprocessing teks yang tepat juga memiliki peran penting dalam meningkatkan akurasi dan efektivitas model machine learning dalam proses klasifikasi teks.

Penggunaan seleksi fitur pada algoritma KNN (k=32) dan Naive Bayes juga memberikan pengaruh yang berbeda. Pada algoritma KNN, seleksi fitur menyebabkan penurunan yang signifikan dalam semua metriks jika dibandingkan tidak menggunakan seleksi fitur. Begitu pula pada algoritma Naive Bayes, seleksi fitur memberikan penurunan performa dalam semua metrik evaluasi untuk semua kombinasi. Walaupun demikian, Kombinasi 3 dan Kombinasi 4 dapat meningkatkan hasil uji secara signifikan, tetapi Kombinasi 2 hanya mendapat peningkatan tipis untuk hasil uji yang telah dilakukan jika dibandingkan dengan Kombinasi 1.

Dapat disimpulkan dalam tabel 4.30 dan gambar 4.9. bahwa performa terbaik untuk seluruh hasil klasifikasi dengan berbagai skenario ditunjukkan

Algoritma Naive Bayes tanpa menggunakan seleksi fitur dan menggunakan kombinasi 4 dengan hasil akurasi 82.0%, precision 82.234%, recall 82.0%, F1-Score 81.967%.

Kombinasi fitur yang terbaik dalam tiap skenario adalah kombinasi 4 yang melibatkan semua fitur preprocessing yang tersedia dalam penelitian ini, terbukti dengan hasil yang telah didapatkan bahwa kombinasi 4 selalu menghasilkan hasil pengujian terbaik di seluruh skenario yang dilakukan, sedangkan kombinasi 1 adalah kombinasi yang menghasilkan hasil pengujian terburuk di hampir seluruh skenario yang dilakukan.

Seleksi Fitur Mutual Information tidak meningkatkan performa pengujian, hal ini dibuktikan dengan nilai pengujian yang cenderung menurun pada seluruh skenario pengujian yang dilakukan, hal ini mungkin disebabkan karena fitur-fitur penting yang banyak tereliminasi dari dataset yang digunakan ketika menggunakan seleksi fitur mutual information.

Dengan demikian, dapat disimpulkan bahwa dalam konteks penelitian ini, penggunaan Algoritma Naive Bayes dengan kombinasi 4 fitur tanpa seleksi fitur mutual information merupakan pendekatan terbaik untuk mencapai hasil klasifikasi yang optimal.

BAB V

PENUTUP

5.1. Kesimpulan

1. Dari hasil pengujian yang telah dilakukan, hasil terbaik pengujian didapatkan menggunakan algoritma Naive Bayes tanpa seleksi fitur mutual information serta menggunakan seluruh fitur preprocessing yang ada dengan hasil akurasi 82.0%, precision 82.234%, recall 82.0%, F1-Score 81.967%. Hal ini membuktikan bahwa algoritma Naive Bayes memiliki performa lebih baik dari K-Nearest Neighbors dalam mengklasifikasi genre anime dengan sinopsis berbahasa inggris.
2. Kombinasi fitur terbaik dalam semua skenario adalah kombinasi 4, yang mencakup semua fitur preprocessing yang tersedia dalam penelitian ini. Hasil pengujian menunjukkan bahwa kombinasi 4 secara konsisten memberikan hasil yang terbaik, sementara kombinasi 1 cenderung memberikan hasil yang terburuk dalam hampir semua skenario.
3. Penerapan Seleksi Fitur Mutual Information tidak menunjukkan peningkatan kinerja pada pengujian, terbukti dengan penurunan nilai pengujian pada berbagai skenario pengujian yang dilakukan.

5.2. Saran

1. Penelitian selanjutnya diharapkan dapat menggunakan dataset yang berbeda sehingga memungkinkan untuk meningkatkan hasil pengujian.

2. Penelitian selanjutnya dapat mencoba algoritma klasifikasi yang berbeda dan menggunakan pendekatan seleksi fitur yang berbeda yang dapat meningkatkan hasil pengujian.

DAFTAR PUSTAKA

PUSTAKA BUKU

Bondebjerg, I. (2015). Film: Genres and Genre Theory. *International Encyclopedia of the Social & Behavioral Sciences: Second Edition, December*, 160–164. https://doi.org/10.1016/B978-0-08-097086-8.95052-9

Cover, T. M., & Thomas, J. A. (2005). Elements of Information Theory. In Elements of Information Theory, John Wiley & Sons, Inc., Hoboken, New Jersey https://doi.org/10.1002/047174882X

Manning, C. D., Raghavan, P., & Schutze, H. (2008). Introduction to Information Retrieval, Cambridge University Press, Cambridge, United Kingdom https://doi.org/10.1017/cbo9780511809071

Nadkarni, P. (2016). *Clinical Research Computing: A Practitioner's Handbook*. Academic Press; 1st edition (May 12, 2016), United States

Prasetyo, E. (2014). Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab. ANDI, Yogyakarta

Rayner, P., Wall, P., & Kruger, S. (2004). AS media studies: The essential introduction. Routledge, London

Utomo, A. P., & Oktora, S. I. (2017). Modul Program Diploma III Sekolah Tinggi Ilmu Statistik METODE REGRESIEdisi Pertama. Sekolah Tinggi Ilmu Statistik.

PUSTAKA MAJALAH, JURNAL ILMIAH ATAU PROSIDING

Ali, H., Salleh, M. N. M., Saedudin, R., Hussain, K., & Mushtaq, M. F. (2019). Imbalance class problems in data mining: A review. *Indonesian Journal of Electrical Engineering and Computer Science*, *14*(3), 1552–1563. https://doi.org/10.11591/ijeecs.v14.i3.pp1552-1563

Aziz, M., & Ong, S. (2023). The Implementation of Japanese Animation (Anime) In Advertising. *Jurnal Indonesia Sosial Sains*, *4*(04), 370–383. https://doi.org/10.59141/jiss.v4i05.810

Akbar, J., Utami, E., & Yaqin, A. (2023). Multi-Label Classification of Film Genres Based on Synopsis Using Support Vector Machine, Logistic Regression and

Naïve Bayes Algorithms. December, 250–255. https://doi.org/10.1109/icitisee57756.2022.10057828

Ariyanti, D., & Iswardani, K. (2020). Teks Mining untuk Klasifikasi Keluhan Masyarakat Pada Pemkot Probolinggo Menggunakan Algoritma Naïve Bayes. Jurnal IKRA-ITH Informatika, 4(3), 125–132.

Arsam, A. (2014). Pembangunan aplikasi video streaming berbasis android di STV Bandung. Jurnal Ilmiah Komputer Dan Informatika (KOMPUTA).

Banlawe, I. A. P., Cruz, J. C. D., Gaspar, J. C. P., & Gutierrez, E. J. I. (2021). Optimal Frequency Characterization of Mango Pulp Weevil Mating Activity using Naïve Bayes Classifier Algorithm. Proceeding - 2021 IEEE 17th International Colloquium on Signal Processing and Its Applications, CSPA 2021, March, 116–120. https://doi.org/10.1109/CSPA52141.2021.9377277

Buslim, N., Oh, L. K., Hardy, M. H. A., & Wijaya, Y. (2022). Comparative Analysis of KNN, Naïve Bayes and SVM Algorithms for Movie Genres Classification Based on Synopsis. Jurnal Teknik Informatika Vol. 15 No. 2, 2022 (169-177).

Chatrina, Siregar, N., Ruli, A, Siregar, R., & Yoga, Distra, Sudirman, M. (2020). Implementasi Metode Naive Bayes Classifier (NBC) Pada Komentar Warga Sekolah Mengenai Pelaksanaan Pembelajaran Jarak Jauh (PJJ). Jurnal Teknologia, 34(1), 102–110. https://aperti.e-journal.id/teknologia/article/view/67

Fathoni, F., Afrianti, E., & Heroza, R. (2020). Klasifikasi Teks dengan Naïve Bayes Classifier (NBC) untuk Pengelompokan Keterangan Laporan dan Durasi Recovery Time Laporan Gangguan Listrik PT. PLN (Persero) WS2JB Area Palembang. JSI: Jurnal Sistem Informasi (E-Journal), 12. https://doi.org/10.36706/jsi.v12i1.9586

Firmanda, R., & Fitriati, D. (2018). Classification of personality type by typology hippocrates - Galenus using hybrid naïve bayes decision tree algorithm. Proceedings of the 3rd International Conference on Informatics and Computing, ICIC 2018, 1–5. https://doi.org/10.1109/IAC.2018.8780447

Laili, A. N., Adikara, P. P., & Adinugroho, S. (2019). Rekomendasi Film Berdasarkan Sinopsis Menggunakan Metode Word2Vec. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *3*(6), 6035–6043. http://j-ptiik.ub.ac.id

Hermansyah, R., & Sarno, R. (2020). Sentiment Analysis about Product and Service Evaluation of PT Telekomunikasi Indonesia Tbk from Tweets Using TextBlob, Naive Bayes & K-NN Method. 2020 International Seminar on Application for

Technology of Information and Communication (ISemantic), 511–516. https://doi.org/10.1109/iSemantic50169.2020.9234238

Jayanti, N. K. T. (2020). Alih Wahana Manga Ao Haru Ride Karya Sakisaka Io ke Dalam Film Live action Karya Sutradara Miki Takahiro. *Jurnal Sakura : Sastra, Bahasa, Kebudayaan Dan Pranata Jepang*, *2*, 63. https://doi.org/10.24843/JS.2020.v02.i02.p01

Park, J., & Lee, D. H. (2020). Parallelly Running k-Nearest Neighbor Classification over Semantically Secure Encrypted Data in Outsourced Environments. IEEE Access, 8, 64617–64633. https://doi.org/10.1109/ACCESS.2020.2984579

Patel, B. A., & Parikh, A. (2020). Impact Analysis of the Complete Blood Count Parameter using Naive Bayes. Proceedings of the 5th International Conference on Inventive Computation Technologies, ICICT 2020, i, 7–12. https://doi.org/10.1109/ICICT48043.2020.9112533

Powers, D. M. W. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. International Journal of Machine Learning Technology p37–63. http://arxiv.org/abs/2010.16061

Rahmayanti, V., Basuki, S., & Hilman, H. (2019). Klasifikasi sinopsis novel menggunakan metode naïve bayes classifier. *Jurnal Repositor*, *1*(2), 125. https://doi.org/10.22219/repositor.v1i2.799

Saputra, A. C., Sitepu, A. B., Stanley, Yohanes Sigit, P. W. P., Sarto Aji Tetuko, P. G., & Nugroho, G. C. (2019). The Classification of the Movie Genre based on Synopsis of the Indonesian Film. Proceeding - 2019 International Conference of Artificial Intelligence and Information Technology, ICAIIT 2019, 201–204. https://doi.org/10.1109/ICAIIT.2019.8834606

Shah, M., Rafi, A., & Perumal, V. (2023). *Investigating Storytelling Differences Between Western and Eastern Computer Animation*. https://doi.org/10.2991/978-2-494069-57-2_15

Sun, H., Wang, Q., Wang, G., Lin, H., Luo, P., Li, J., Zeng, S., Xu, X., & Ren, L. (2018). Optimizing kNN for mapping vegetation cover of arid and semi-arid areas using landsat images. Remote Sensing, 10(8). https://doi.org/10.3390/rs10081248

PUSTAKA LAPORAN PENELITIAN

Saleh, A. (2015). Klasifikasi Gejala Depresi Pada Manusia dengan Metode Naïve Bayes Menggunakan Java. Skripsi Tesis, STMIK AKAKOM, Yogyakarta.