



Tugas Besar Analisis Big Data
Kelompok 9 RB

ANALISIS SENTIMEN TERHADAP ULASAN GAME PUBG DALAM DATA STEAM TAHUN 2021: PENDEKATAN BIG DATA DENGAN PYSPARK

Fakultas Sains
Program Studi Sains Data
Institut Teknologi Sumatera
2024

Anggota Kelompok 9 RB:



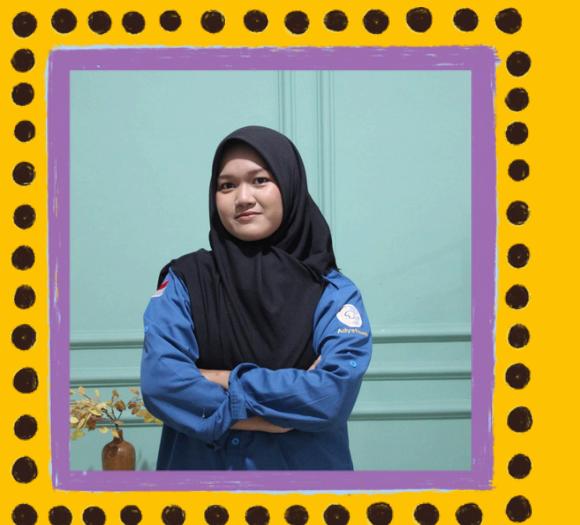
Evan Aprianto
121450024



M. Gilang Martiansyah M.
121450056



Meinisa
121450076



Syifa Firnanda
121450094



Dara cantika Dewi
121450127



Pendahuluan

Latar Belakang

Teknologi hiburan, khususnya dalam game, mengalami perkembangan pesat dengan banyaknya platform online seperti Steam. Steam menjadi distributor terkemuka dengan jutaan pengguna aktif dan ribuan game. Ulasan game di Steam menjadi penting untuk mengetahui kualitasnya, dan analisis sentimen digunakan untuk mengklasifikasikan ulasan. Dalam penelitian ini, Apache Spark digunakan untuk mengatasi volume data besar dengan efisien. Fokus analisis sentimen adalah pada game PUBG di Steam, dengan metode VADER untuk membuat label sentimen. Label ini digunakan dalam model klasifikasi Naive Bayes dan Regresi Logistik dalam PySpark untuk mengkategorikan persepsi pemain terhadap PUBG dan mengevaluasi akurasinya.

Pendahuluan



Rumusan Masalah

1. Bagaimana perbandingan persepsi pemain berdasarkan distribusi sentimen (positif, negatif, dan netral) dalam ulasan game PUBG di Steam tahun 2021 dengan metode VADER?
2. Bagaimana akurasi dari penerapan PySpark dan VADER dalam analisis sentimen ulasan game PUBG menggunakan model Naive Bayes dan Regresi Logistik?

Tujuan Penelitian

1. Menganalisis hasil klasterisasi persepsi pemain pemain berdasarkan distribusi sentimen (positif, negatif, dan netral) dalam ulasan terhadap game PUBG di Steam tahun 2021 dengan metode VADER.
2. Mengevaluasi efektivitas penggunaan model Naive Bayes dan Regresi Logistik dalam PySpark untuk analisis sentimen ulasan game di Steam berdasarkan hasil akurasi.

Batasan Masalah

1. Penelitian ini hanya fokus pada ulasan game PUBG di Steam tahun 2021 dan hanya akan menggunakan ulasan dalam bahasa Inggris.
2. Analisis sentimen dilakukan dengan metode VADER sebagai membuat label distribusi sentimen (positif, negatif, dan netral), serta model Naive Bayes dan Regresi Logistik untuk membuat prediksi mengenai kelompok sentimen yang paling mungkin.

Landasan Teori

Analisis Sentimen

Analisis Sentimen adalah proses mengelompokkan polaritas dari teks dalam suatu dokumen, untuk menentukan apakah pendapat yang diungkapkan bersifat positif, negatif, atau netral.

VADER

Valence Aware Dictionary and Sentiment Reasoner (VADER) adalah metode analisis sentimen yang dapat mengukur intensitas emosional berdasarkan kamus Lexicon yang tersedia.

Naive Bayes

Naive Bayes merupakan algoritma yang digunakan untuk menghitung probabilitas suatu kejadian berdasarkan informasi yang sudah diketahui sebelumnya.

Regresi Logistik

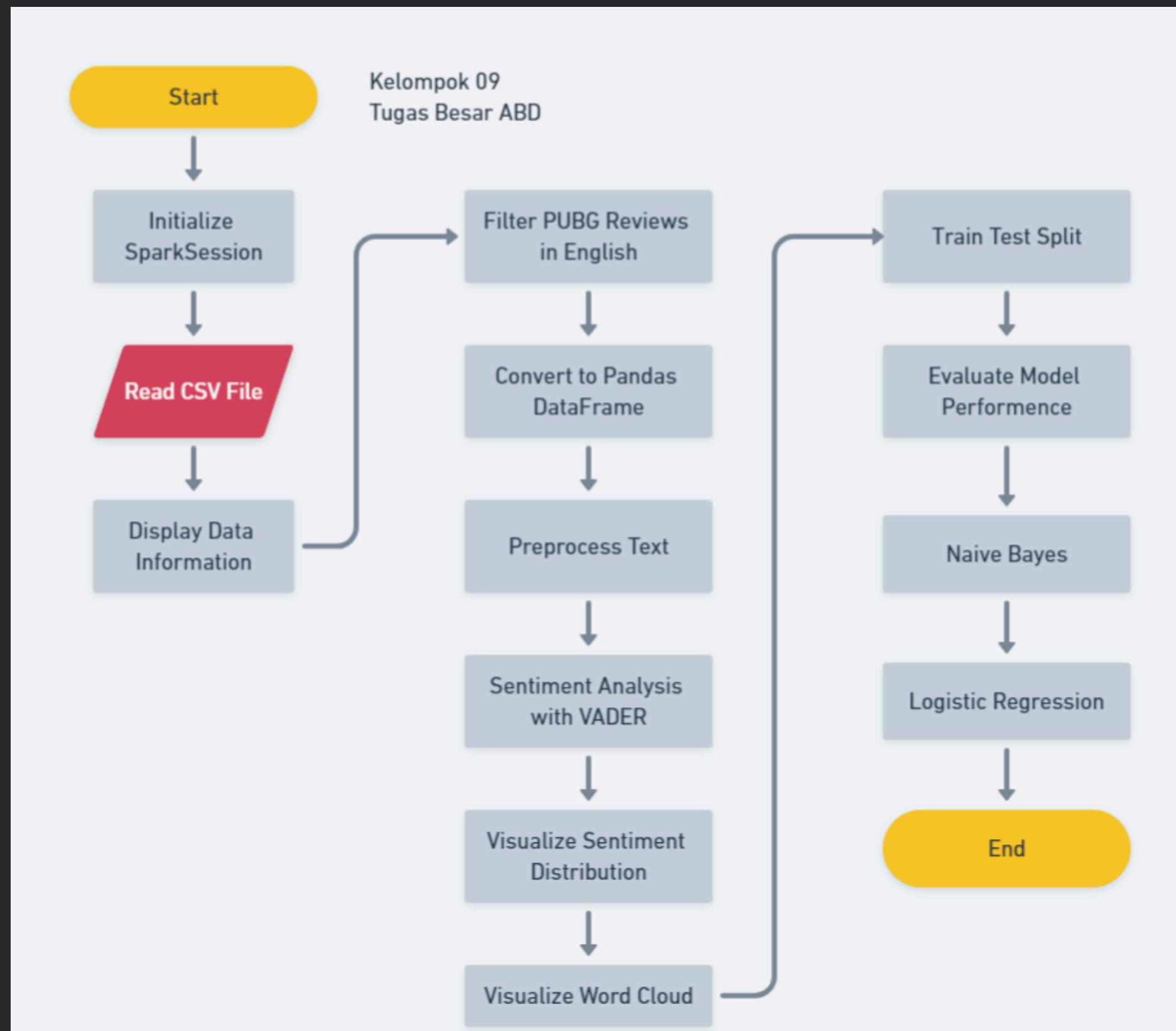
Regresi Logistik merupakan algoritma yang digunakan untuk memprediksi hasil dari suatu variabel berdasarkan satu atau lebih variabel independen.

Apache Spark

Apache Spark adalah sebuah framework pemrosesan data terdistribusi yang dikembangkan oleh Apache Software Foundation. Spark dirancang untuk memproses data besar dengan menggunakan teknologi caching dalam memori dan eksekusi kueri yang dioptimalkan.



Desain Penelitian



Metode Penelitian

Data Collection and Preparation

Dataset ulasan PUBG tahun 2021 diunduh dari Steam, mencakup 40.848.659 baris dan 23 kolom, lalu dimuat ke PySpark untuk pemrosesan efisien. Data dieksplorasi untuk memahami struktur dan distribusi sentimen, kemudian dibersihkan dari duplikat dan nilai hilang. Setelah pembersihan, dataset siap untuk analisis sentimen.

Sentiment Analysis

Analisis sentimen dilakukan menggunakan library VADER, yang memiliki kamus dan aturan untuk menentukan sentimen positif, negatif, atau netral dari setiap ulasan.



Text Processing

Setelah memastikan kualitas data yang baik, langkah berikutnya adalah mempersiapkan teks mentah untuk analisis sentimen. Pertama, ulasan di-tokenisasi menjadi kata-kata individual. Selanjutnya, semua token diubah menjadi huruf kecil. Teks dinormalisasi dengan menghapus tanda baca dan karakter khusus. Kata-kata yang tidak penting dihapus untuk meningkatkan kualitas teks. Akhirnya, kata-kata diubah menjadi bentuk dasar atau dipangkas untuk konsistensi data.

Metode Penelitian

Model Training

Dalam model training, dataset dibagi menjadi 80% untuk pelatihan dan 20% untuk pengujian. Fitur diekstraksi dari teks ulasan menggunakan metode TF-IDF, dan model klasifikasi sentimen (seperti Naive Bayes dan Logistic Regression) dilatih dengan data pelatihan. Ini memungkinkan model memahami hubungan antara fitur numerik dan label sentimen untuk mengklasifikasikan sentimen ulasan dengan akurasi baik.

Model Performance Evaluation

Untuk mengevaluasi kinerja model, berbagai matrik evaluasi seperti akurasi, presisi, recall, dan F1-score digunakan. Confusion matrix juga dibuat untuk mengevaluasi klasifikasi model pada set pelatihan dan pengujian, yang membantu dalam memahami seberapa baik model dapat mengklasifikasikan label sentimen dengan benar. Selain itu, dilakukan validasi silang untuk memastikan konsistensi kinerja model dan mencegah overfitting.



Hasil Pembahasan

Analisis sentimen terhadap ulasan game PUBG menggunakan dua model, yaitu Naive Bayes dan Logistic Regression. Data menunjukkan bahwa ulasan positif lebih banyak dibandingkan negatif dan netral, yang mengindikasikan kepuasan pengguna terhadap game.

Dalam evaluasi model, Logistic Regression menunjukkan performa yang lebih baik dengan akurasi 94%, sedangkan Naive Bayes hanya 75%. Logistic Regression terbukti lebih efektif dalam mengklasifikasikan ulasan dengan kesalahan yang lebih rendah dibanding Naive Bayes. Model Naive Bayes mengalami kesulitan dalam membedakan antara kelas "Negatif" dan kelas lainnya, sementara Logistic Regression dapat membedakan ketiga kelas dengan lebih baik.



Hasil Pembahasan

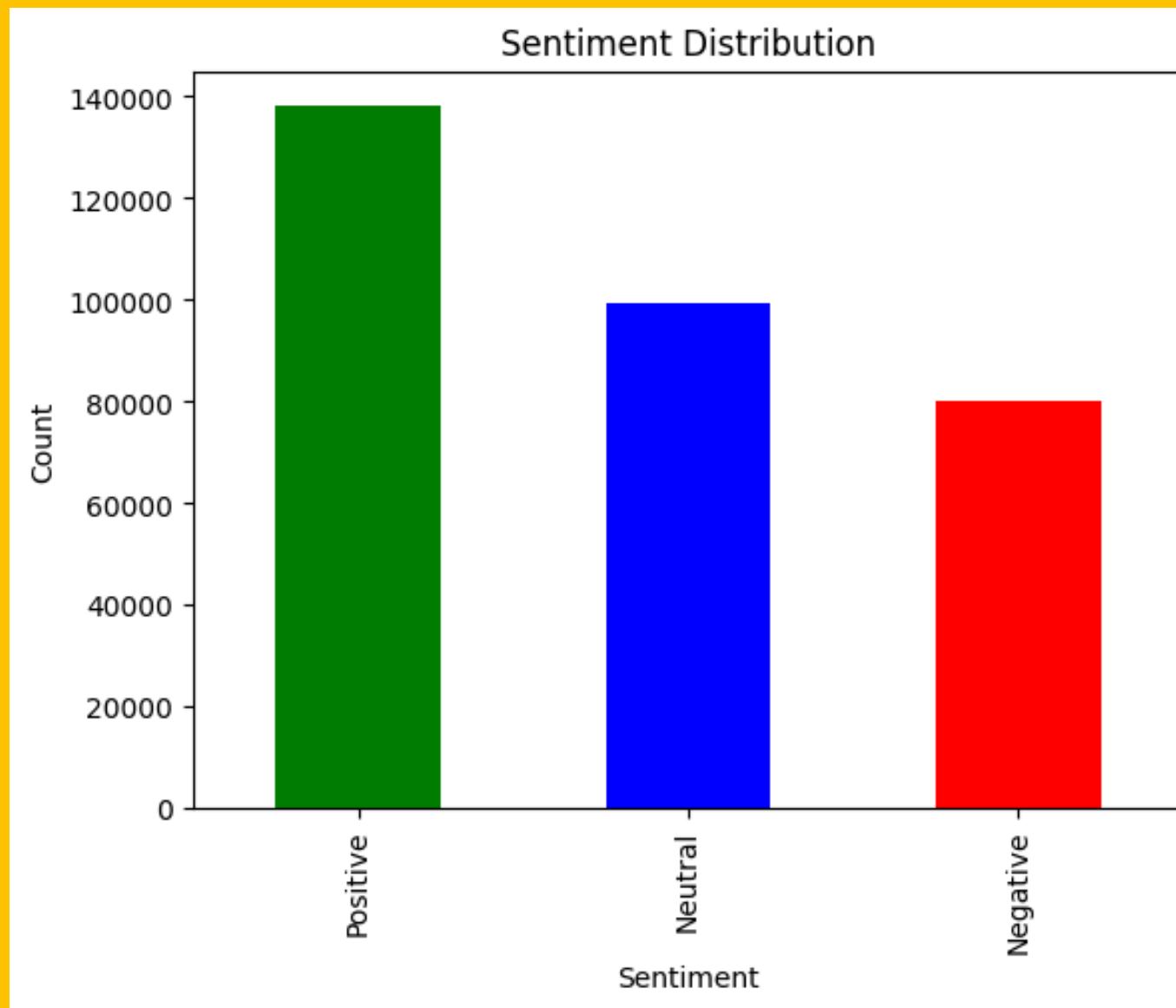
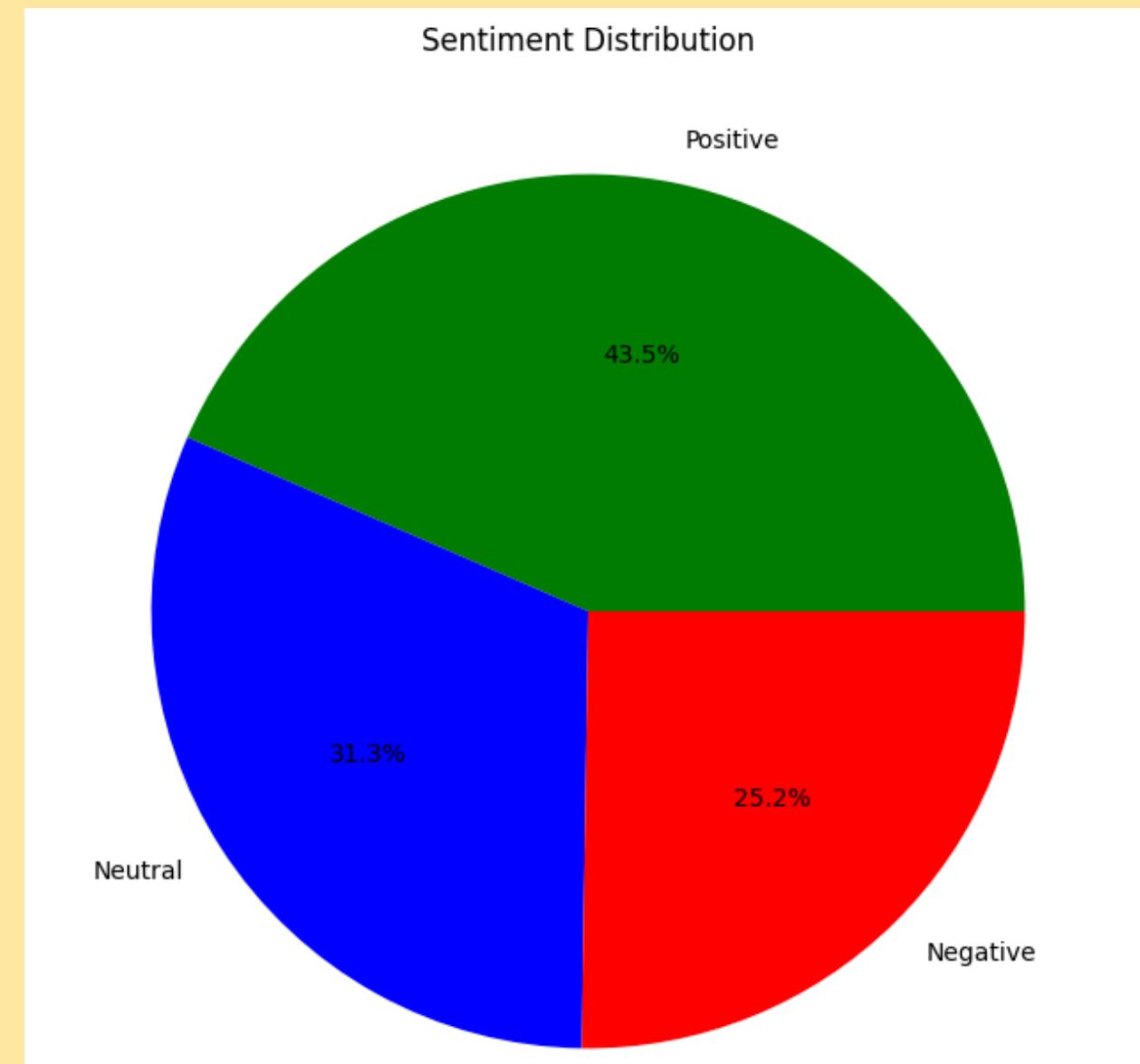


Diagram Batang Distribusi Sentimen

Pada diagram Barchart didapatkan bahwa ulasan positif mendominasi dengan sekitar 139.000 kata, diikuti oleh ulasan netral (99.000 kata) dan negatif (79.000 kata), menunjukkan sentimen yang umumnya positif dari pengguna



Pie Chart Distribusi Sentimen

Pada diagram Pie Chart didapatkan bahwa persentase kata positif adalah 43,5%, netral 31,3%, dan negatif 25,2%. Ini menegaskan bahwa mayoritas ulasan bersifat positif, namun masih ada sejumlah ulasan netral dan negatif yang perlu diperhatikan untuk perbaikan lebih lanjut.

Word Cloud

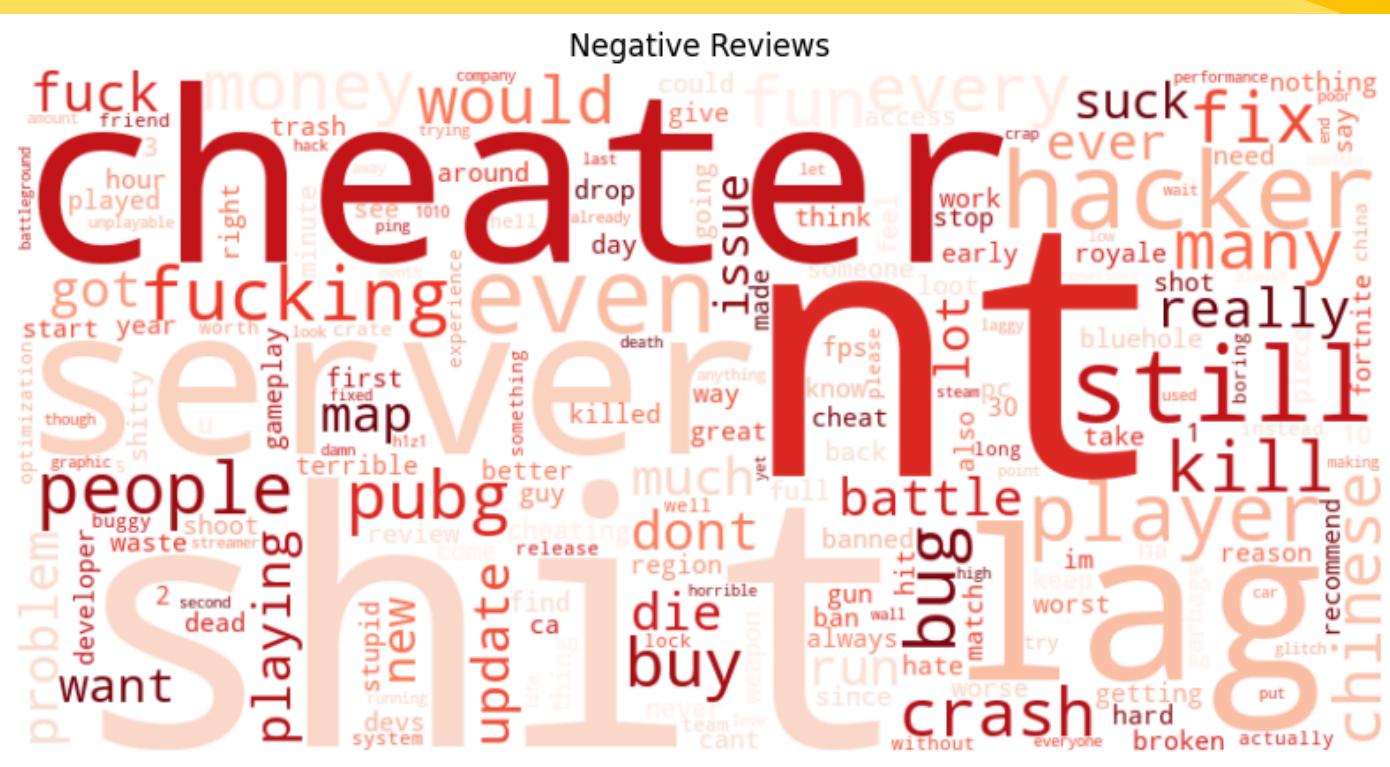
Word Cloud Positif: Kata-kata seperti "Best", "Great", "Better", "Fun", dan "Friend" paling sering muncul, menunjukkan bahwa pemain menghargai pengalaman bermain yang menyenangkan, peningkatan kualitas, dan aspek sosial dalam game PUBG.

Word Cloud Netral: Kata-kata dalam ulasan netral memiliki proporsi yang seragam, mencerminkan pandangan yang lebih moderat dan seimbang, serta mencakup baik elemen positif maupun negatif yang tidak terlalu kuat.

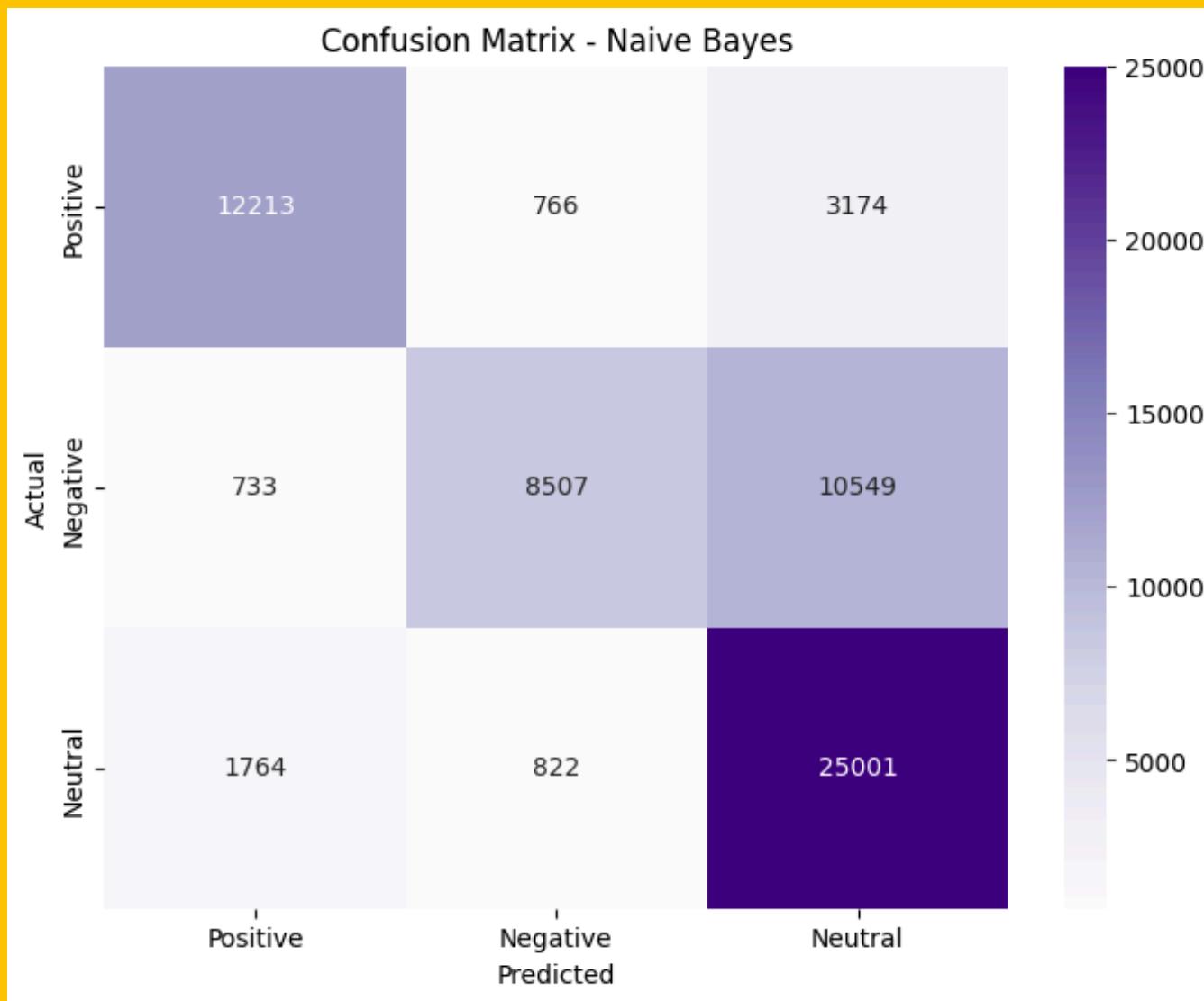
Word Cloud

Negatif: Kata-kata seperti "Cheater", "Lag", "Bug", dan "Crash" sering muncul, yang berarti menyoroti

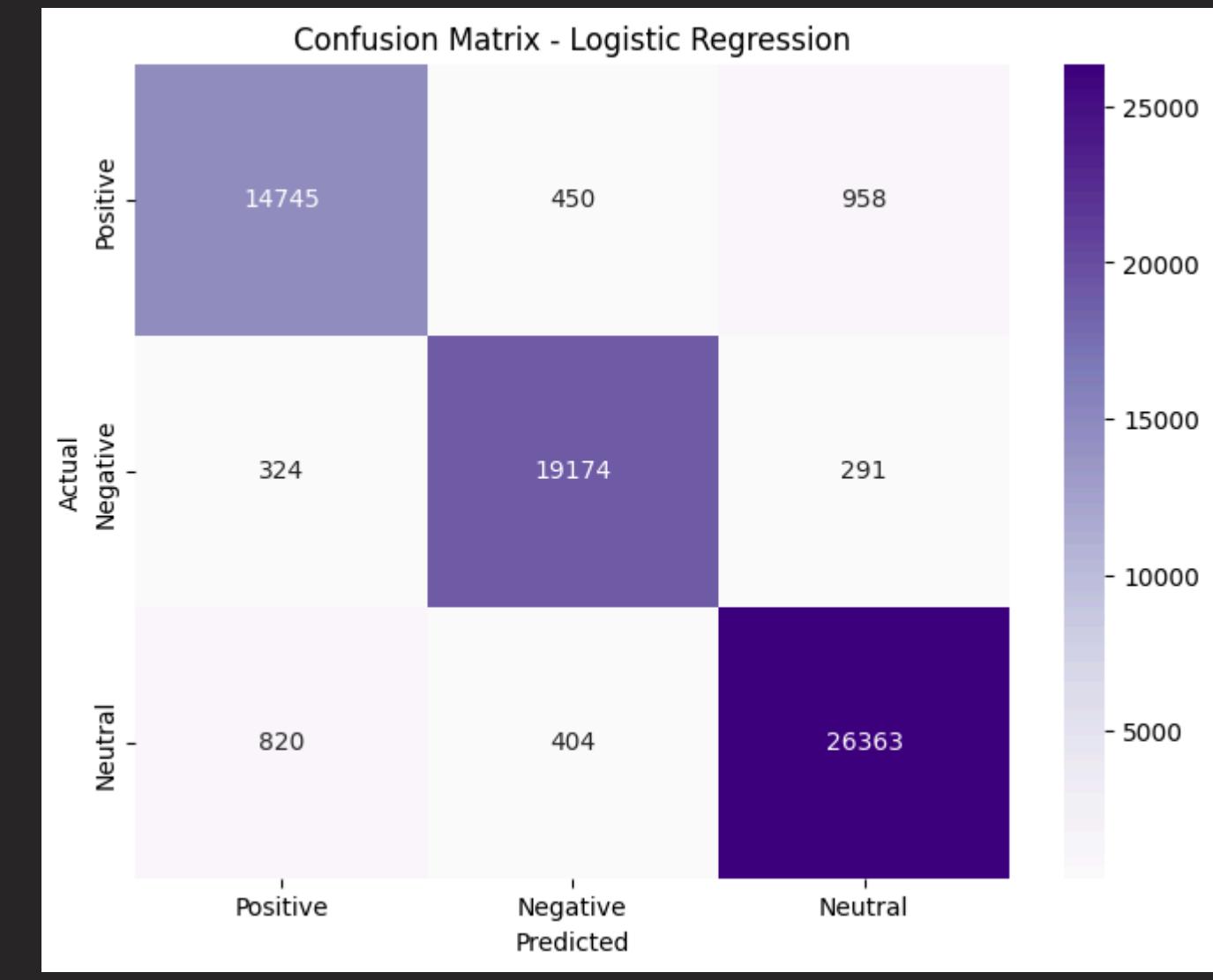
masalah teknis dan perilaku curang yang menyebabkan ketidakpuasan di kalangan pemain.



Confusion Matrix



Pada model Naive Bayes mengklasifikasikan kelas "Positif" dengan kategori "True Positif" dengan jumlah 12.213 kata. Kelas "Negatif" dengan kategori "True Negatif" dengan jumlah 8.507 kata. Dan kelas "Netral" dengan kategori "True Netral" dengan jumlah 25.001 kata.



Pada model Logistic Regression mengklasifikasikan kelas "Positif" dengan kategori "True Positif" dengan jumlah 14.745 kata. Kelas "Negatif" dengan kategori "True Negatif" dengan jumlah 19.174 kata. Dan kelas "Netral" dengan kategori "True Netral" dengan jumlah 26.363 kata.

KESIMPULAN



Dari hasil visualisasi data yang dihasilkan, mayoritas ulasan muncul dengan kata-kata bersifat “Positif” (138129 kata) dibandingkan “Netral” (99379 kata) dan “Negatif” (80135 kata). Selain itu, proporsi ulasan “Netral” dan “Negatif” yang signifikan juga menunjukkan bahwa terdapat ruang untuk peningkatan lebih lanjut dalam pengalaman pengguna. Pengembang aplikasi dapat menggunakan informasi ini untuk fokus pada aspek-aspek yang paling dihargai oleh pengguna, serta mengidentifikasi dan memperbaiki area yang menimbulkan ketidakpuasan.

Logistic Regression menunjukkan performa yang lebih baik dalam klasifikasi teks dibandingkan Naive Bayes, dengan akurasi yang lebih tinggi dan kemampuan yang lebih baik dalam mengurangi kesalahan klasifikasi antara kelas yang berbeda. Oleh karena itu, penggunaan model Logistic Regression disarankan untuk analisis sentimen ulasan game PUBG di Steam agar dapat memberikan hasil yang lebih akurat dan efektif. Pengembang game dapat menggunakan hasil analisis ini untuk meningkatkan kualitas dan kepuasan pengguna terhadap game PUBG.



THANK YOU

Sekian presentasi dari kami, kami pamit undur diri
karena kalau maju saingannya satu prodi



