Tugas Besar Analisis Big Data

Accepted 23 Mei 2024

Open Access

Penerapan Spark dengan Metode *Naive Bayes Classifier* Pada Analisis Sentimen Tingkat Kepuasan Pengguna Aplikasi Go-Jek di *Google Play Store*

Gede Moena_121450014 1[*a], Hanna Sajidah_121450060 2 [b], Ari Sigit_121450069 3 [c,] Elsyah Sapyrah_121450096 4[d,] Tarisah_121450141 5[e]*

Corresponding E-mail:

gede.121450014@student.itera.ac.id ;

ari.121450069@student.itera.ac.id;

hanna.121450060@student.itera.ac.id;

elsyah.121450096@student.itera.ac.id;

tarisah.121450141@student.itera.ac.id;

Abstrak

In the current digital era, mobile applications have become an inseparable part of everyday life. One application that is widely used by Indonesians is GO-JEK, a super application that provides various services ranging from transportation to food delivery. User satisfaction with these services is often expressed through ratings and reviews on the Google Play Store, which is a valuable source of data for understanding user perceptions and experiences. In this research, review and rating data from the Google Play Store will be collected and analyzed using the Naive Bayes method processed with Apache Spark. This analysis process will include data collection, data analysis preprocessing, training models, and evaluating results. From the results of the analysis that has been carried out, an accuracy value of 0.8868923515871916 is obtained, meaning that the Naive Bayes model can explain the classification of user reviews into each class by 88%.

Kata kunci: *GO-JEK, Review, Naive Bayes, Apache Spark, Classification.*

Abstrak

Pada era digital saat ini, aplikasi mobile telah menjadi bagian tak terpisahkan dari kehidupan sehari-hari. Salah satu aplikasi yang banyak digunakan oleh masyarakat Indonesia adalah GO-JEK, sebuah aplikasi super yang menyediakan berbagai layanan mulai dari transportasi hingga pengantaran makanan. Kepuasan pengguna terhadap layanan ini sering kali diekspresikan melalui rating dan review di Google Play Store, yang menjadi sumber data berharga untuk memahami persepsi dan pengalaman pengguna. Dalam penelitian ini, data ulasan dan rating dari Google Play Store akan dikumpulkan dan dianalisis menggunakan metode Naive Bayes yang diproses dengan Apache Spark. Proses analisis ini akan mencakup pengumpulan data, praproses data, pelatihan model, dan evaluasi

hasil analisis. Dari hasil analisis yang telah dilakukan diperoleh nilai akurasi sebesar 0.8868923515871916, artinya model Naive Bayes dapat menjelaskan klasifikasi dari ulasan pengguna ke dalam setiap kelas sebesar 88%.

Kata Kunci : GO-JEK, Review, Naive Bayes, Apache Spark, Klasifikasi.

Pendahuluan

Pada era digital saat ini, aplikasi mobile telah menjadi bagian tak terpisahkan dari kehidupan sehari-hari. Salah satu aplikasi yang banyak digunakan oleh masyarakat Indonesia adalah *GO-JEK*, sebuah aplikasi yang menyediakan berbagai layanan mulai dari transportasi hingga pengantaran makanan. Kepuasan pengguna terhadap layanan ini sering kali diekspresikan melalui *rating* dan *review* di *Google Play Store* yang menjadi sumber data berharga untuk memahami persepsi dan pengalaman pengguna.

Analisis sentimen adalah proses menganalisis teks digital untuk menentukan apakah nada emosional pesan tersebut positif, negatif, atau netral [1]. Salah satu metode yang populer untuk melakukan analisis sentimen adalah *Naive Bayes* yang merupakan salah satu metode dalam *supervised learning* yang biasanya digunakan untuk klasifikasi. *Naive Bayes* adalah algoritma probabilistik yang didasarkan pada Teorema Bayes dengan asumsi bahwa fitur-fitur dalam data bersifat independen satu sama lain [2]. Algoritma ini sederhana namun sangat efektif dan cepat dalam melakukan klasifikasi terutama pada data teks.

Dalam konteks analisis sentimen terhadap ulasan GO-JEK di *Google Play Store*, Apache Spark memainkan peran penting sebagai kerangka kerja komputasi terdistribusi yang memungkinkan pemrosesan data besar dengan kecepatan tinggi. *Apache Spark* adalah sistem pemrosesan terdistribusi sumber terbuka yang digunakan untuk beban kerja big data[3]. Dengan kemampuannya dalam pemrosesan data terdistribusi, Spark memungkinkan pengolahan jutaan ulasan dan rating dari pengguna dengan efisien, sehingga memungkinkan analisis sentimen yang mendalam dan responsif terhadap persepsi pengguna terhadap berbagai aspek layanan yang ditawarkan oleh GO-JEK. Selain itu, fitur in-memory computing dari Spark mempercepat operasi data yang krusial dalam menangani volume data besar yang dihasilkan oleh aplikasi populer seperti GO-JEK.

Dalam penelitian ini, data ulasan dan rating dari Google Play Store akan dikumpulkan dan dianalisis menggunakan metode Naive Bayes yang diproses dengan Apache Spark . Proses analisis ini akan mencakup pengumpulan data, praproses data, pelatihan model, dan evaluasi hasil analisis.

Penelitian ini bertujuan untuk menerapkan Spark dan metode Naive Bayes pada analisis sentimen ulasan aplikasi GO-JEK berdasarkan rating dan review yang ada di Google Play Store. Melalui penelitian ini diharapkan dapat diperoleh wawasan yang lebih mendalam mengenai sentimen pengguna terhadap berbagai aspek layanan yang ditawarkan oleh GO-JEK. Selain itu bertujuan untuk mengevaluasi kinerja Apache Spark dalam menangani pemrosesan data besar serta efektivitas Naive Bayes dalam klasifikasi sentimen teks.

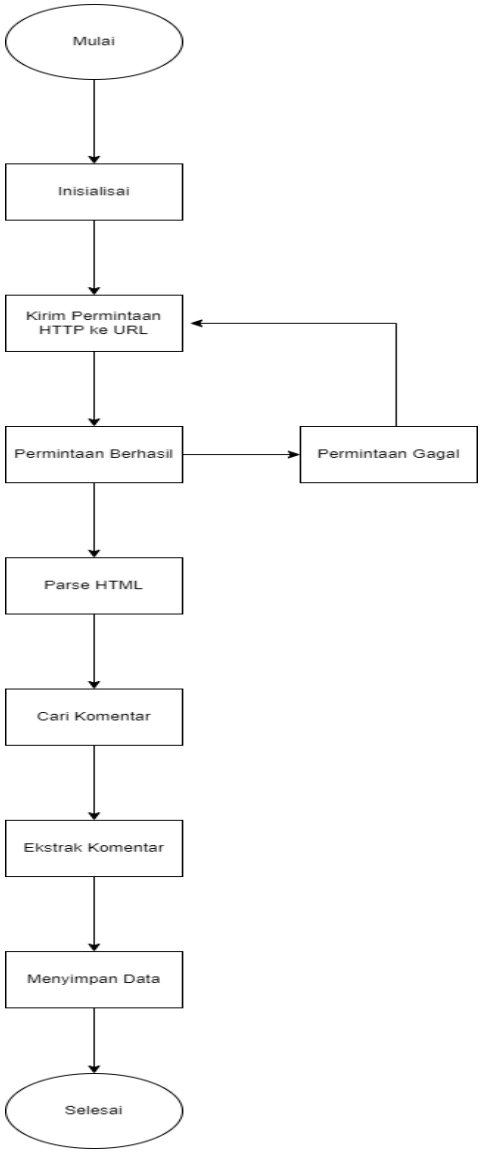
Metode

Pengumpulan Data

1. *Data Mining*

Data Mining merupakan proses pengumpulan dan pengolahan data yang bertujuan untuk mengekstrak informasi penting pada data sehingga kita mendapat knowledge dari data tersebut[4]. Pada Penelitian ini menggunakan metode *web scraper* dimana *web scraper* merupakan teknik ekstraksi data dalam jumlah besar dari berbagai website secara otomatis, web scraper ini digunakan untuk melakukan ekstraksi data yang ada pada kolom komentar pada aplikasi Go-Jek di google play store dengan tujuan melihat tingkat kepuasan pengguna pada aplikasi tersebut dengan menggunakan Analisis sentimen.

flowchart ekstraksi data



Gambar 1. Alur Ekstraksi Data

2. Batch Pipelines

Pada penelitian ini juga menerapkan konsep batch pipelines yang merupakan sebuah proses pemrosesan data yang dilakukan secara bertahap dalam batch. Dalam pendekatan ini, data ulasan diproses dalam interval waktu tertentu atau ketika terdapat kumpulan data yang mencukupi. Penggunaan batch pipelines lebih sesuai ketika data bersifat statis atau tidak berubah dengan cepat, seperti data ulasan yang mungkin tidak perlu dianalisis secara real-time.

Preprocessing data

1. *Data cleaning*

Data cleaning merupakan proses memperbaiki atau menghapus kesalahan, ketidakkonsistenan, dan ketidakakuratan dalam kumpulan data [5]. Tujuannya adalah meningkatkan kualitas data dan memastikan bahwa data tersebut dapat diandalkan untuk analisis. Pembersihan data melibatkan berbagai langkah-langkah untuk menangani kesalahan data, duplikasi, nilai yang hilang, serta inkonsistensi dalam data. Data cleaning sangat penting untuk dilakukan karena,

1. Meningkatkan kualitas data : Data yang lebih akurat dan dapat diandalkan, sehingga menghasilkan analisis yang valid.
2. Efisensi Proses : Data yang bersih mempermudah dan mempercepat proses analisis dan pengambilan keputusan
3. Menghindari Kesalahan : Data yang kotor dapat menghasilkan kesalahan dalam analisis, yang dapat berakibat pada keputusan yang salah.

2. *Manipulation Data*

Manipulation Data adalah proses memanipulasi atau mengubah informasi agar lebih terorganisir dan mudah dipahami, serta mempersiapkannya untuk analisis lebih lanjut [6]. Manipulasi data bisa melibatkan berbagai operasi, mulai dari penambahan atau penghapusan kolom, penggabungan dataset, perubahan format data, hingga perhitungan agregat, dengan menggunakan berbagai teknik manipulasi data, dapat mempersiapkan dan mengelola data untuk analisis lebih lanjut.

Processing Data

1. Labelling

Dalam proses analisis sentimen, langkah awal adalah memberi label pada data. Setiap ulasan diberi label sentimen positif, negatif, atau netral berdasarkan

Diterbitkan oleh: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera , Lampung Selatan, Indonesia

konten ulasan tersebut. Proses pemberian label ini memungkinkan model untuk belajar dan mengenali pola-pola tertentu yang terkait dengan setiap kategori sentimen. Dengan label yang tepat, model *Naive Bayes* Classifier dapat melatih dirinya untuk mengklasifikasikan sentimen dengan akurasi yang lebih baik serta membantu dalam menganalisis tingkat kepuasan pengguna aplikasi Go-Jek secara lebih mendalam.

2. *Case Folding*

Case folding adalah proses mengubah semua huruf dalam teks menjadi huruf kecil (lowercase) sebelum dilakukan tahapan selanjutnya dalam analisis teks [7]. Tujuan utama dari case folding adalah untuk menghindari perbedaan hasil akibat variasi penulisan huruf besar dan kecil dalam suatu teks.

Misalnya, dalam analisis sentimen ulasan pengguna aplikasi Go-Jek kata "Baik" dan "baik" harus dianggap sama agar tidak menghasilkan klasifikasi yang berbeda. Dengan menerapkan case folding, kedua kata tersebut akan diubah menjadi "baik", sehingga tidak ada perbedaan antara keduanya dalam analisis.

3. *Tokenizer*

Tokenization atau tokenisasi yaitu tahap untuk pemotongan pada string input sesuai kata penyusunnya. Proses tokenisasi ini akan mempermudah dalam membedakan karakter yang dapat dijadikan sebagai pemisah kata.

Tokenizer untuk analisis sentimen komentar aplikasi *Gojek di Play Store* merupakan alat yang digunakan untuk memisahkan kata-kata atau frasa-frasa dalam komentar pengguna menjadi token-token yang lebih kecil. Dengan menggunakan tokenizer, komentar-komentar tersebut dapat dipecah menjadi token-token seperti kata-kata, frasa, atau tanda baca, yang kemudian dapat diolah lebih lanjut menggunakan teknik-teknik analisis sentimen seperti klasifikasi teks atau analisis pola kata-kata untuk menentukan sentimen umum dari setiap komentar [8].

4. *Stop word removal*

Stopword Removal merupakan bagian dari tahapan preprocessing teks yang bertujuan untuk menghapus kata yang tidak relevan di dalam suatu kalimat berdasarkan daftar stopword [9]. Daftar stop words berisi kata-kata umum seperti "nya", "dan", "di", "ada", dan sebagainya yang sering muncul dalam teks namun tidak memberikan informasi yang berguna dalam analisis sentimen. Dengan menghilangkan stop words, kita dapat meningkatkan akurasi model dan fokus pada kata-kata yang lebih bermakna dalam menentukan sentimen.

5. *Count vectorizer*

CountVectorizer adalah metode dasar pengukuran kata dengan menghitung jumlah kemunculan masing-masing kata pada dokumen, sehingga dapat disebut juga sebagai metode raw count [10]. Teknik ini mengubah teks menjadi representasi numerik yang dapat dimengerti oleh model pemrosesan bahasa alami.

6. *Pipeline dan transformer*

Pipeline adalah rangkaian langkah-langkah pemrosesan data yang dijalankan secara berurutan. Dalam konteks analisis sentimen menggunakan Naive Bayes Classifier pada Apache Spark, pipeline digunakan untuk mengatur alur kerja dari awal hingga akhir, termasuk pra-pemrosesan teks, pelatihan model, dan evaluasi performa.

Transformer adalah komponen utama dalam pipeline yang mengubah DataFrame dari satu bentuk ke bentuk lainnya melalui serangkaian operasi. Dalam proses analisis sentimen, transformer digunakan untuk melakukan pra-pemrosesan teks, seperti tokenisasi, penghapusan stop words, ekstraksi fitur, dan lain-lain. Contohnya, tokenisator mengubah ulasan teks menjadi kumpulan token kata, sedangkan penghapus stop words menghilangkan kata-kata yang tidak relevan dari teks. Transformer membantu mempersiapkan data mentah menjadi format yang dapat digunakan oleh model untuk pelatihan dan evaluasi. Dengan menggunakan transformer dalam pipeline, alur kerja analisis sentimen menjadi lebih terstruktur dan efisien.

7. *TF-IDF*

Skema TF-IDF (Term Frequency-Inverse Document Frequency) adalah teknik yang digunakan dalam pemrosesan bahasa alami dan pengambilan informasi untuk mengevaluasi pentingnya suatu kata

dalam dokumen teks relatif terhadap kumpulan dokumen yang lebih besar [11].

1. *Term Frequency (TF)*

Mengukur seberapa sering sebuah kata muncul dalam sebuah dokumen. Formula yang umum digunakan adalah TF = (jumlah kemunculan kata dalam dokumen) / (jumlah kata total dalam dokumen). Misalnya, jika kata "Gojek" muncul 5 kali dalam sebuah komentar yang memiliki total 100 kata, maka TF untuk kata "Gojek" dalam dokumen tersebut adalah 5/100 = 0.05.

2. *Inverse Document Frequency (IDF)*

Mengukur seberapa umum sebuah kata di seluruh kumpulan dokumen. Kata-kata yang sering muncul di banyak dokumen dianggap kurang penting dalam membedakan dokumen satu dengan yang lain. IDF dihitung dengan membagi jumlah total dokumen dalam kumpulan dokumen dengan jumlah dokumen yang mengandung kata tersebut. Hasilnya kemudian di-logaritmakan. Hal ini menghasilkan skor yang lebih besar untuk kata-kata yang jarang muncul di seluruh kumpulan dokumen.

Visualisasi Word Cloud

Word cloud merupakan visualisasi dari kumpulan kata yang sering disebut dalam sebuah media tertentu[12]. Teknik ini digunakan dalam analisis data teks untuk menampilkan kata-kata yang sering muncul dalam sebuah kumpulan teks dengan cara yang menarik dan mudah dipahami. Fungsi utama dari wordcloud adalah untuk mengidentifikasi kata kunci dan istilah penting dalam teks dengan cepat. Kata-kata yang lebih sering muncul akan ditampilkan dalam ukuran yang lebih besar, memudahkan identifikasi pola dan tren dalam data. Selain itu, word cloud juga bermanfaat dalam analisis sentimen, memberikan gambaran umum tentang sentimen teks melalui kata-kata yang sering digunakan.

Modelling

1. Splitting data (Data train dan data test)

Pembagian data adalah langkah kritis dalam proses pemodelan machine learning yang bertujuan untuk menguji kinerja model secara objektif. Dalam konteks analisis sentimen terhadap ulasan GO-JEK di Google Play Store, data perlu dibagi menjadi dua subset utama: data pelatihan (train) dan data pengujian (test).

a. Data Pelatihan (*Train Data*)

Data pelatihan, atau sering disebut sebagai data train adalah bagian dari dataset yang digunakan untuk melatih model machine learning. Model akan belajar dari data ini untuk menemukan pola dan hubungan antara fitur dan label (dalam kasus ini, rating bintang dan teks ulasan). Data train harus mewakili variasi yang ada dalam dataset untuk memastikan model dapat memahami pola yang umum dan dapat diterapkan dengan baik pada data baru.

b. Data Pengujian (*Test Data*)

Data pengujian, atau sering disebut sebagai data test, adalah bagian dari dataset yang tidak digunakan selama proses pelatihan model, tetapi digunakan untuk menguji kinerja model yang telah dilatih. Data test digunakan untuk mengevaluasi seberapa baik model dapat melakukan prediksi terhadap data yang belum pernah dilihat sebelumnya. Pembagian data test harus mencerminkan distribusi yang sama dengan data train untuk menghasilkan evaluasi yang adil.

2. *Naive Bayes*

Naive Bayes adalah algoritma pembelajaran mesin populer yang digunakan untuk tugas klasifikasi. Hal Itu didasarkan pada teorema Bayes dan mengasumsikan bahwa fitur-fiturnya tidak tergantung satu sama lain. Pengklasifikasi Naive Bayes menghitung probabilitas titik data milik kelas tertentu berdasarkan probabilitas fitur yang diberikan kelas itu. Kelas dengan probabilitas tertinggi kemudian ditugaskan ke titik data. Pengklasifikasi Naive Bayes banyak digunakan dalamklasifikasi teks, pemfilteran spam, dan analisis sentimen, di antara aplikasi lainnya

$$P(c|x) = P(x|c) * P(c)/P(x)$$

Diterbitkan oleh: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera , Lampung Selatan, Indonesia

Dimana :

P(c|x) = probabilitas kelas c diberikan data x

P(x|c) = probabilitas kelas x diberikan data c

P(c) = probabilitas kelas c

P(x) = banyaknya pasangan rating asli dan rating prediksi

Evaluasi Model:

1. K-fold Cross Validation

   K-Fold Cross Validation adalah proses pengujian dengan melipat (fold) data menjadi bagian set data dengan jumlah yang sama sebanyak K lipatan/pecahan dan menguji model pada setiap subset. Ini membantu dalam mengurangi bias dan overfitting yang bisa terjadi jika hanya menggunakan satu subset data untuk validasi [14].

   Ada berbagai variasi dalam K-Fold, seperti validasi silang k-fold berlapis, yang memastikan bahwa setiap fold berisi representasi proporsional dari kelas yang berbeda dalam kumpulan data. Pilihan k tergantung pada ukuran dataset dan tingkat presisi yang diinginkan dalam estimasi performa model.

   K-fold cross validation membantu mengatasi masalah overfitting dan underfitting pada model, serta memungkinkan penggunaan data yang lebih efisien. Teknik ini juga memungkinkan pengguna untuk mengevaluasi model dengan lebih akurat daripada

   teknik validasi model lainnya[15].

2. Confussion Matriks

   Confusion Matrix adalah alat evaluasi yang digunakan dalam pembelajaran mesin dan statistik untuk menilai kinerja model klasifikasi. Matriks ini memberikan gambaran lengkap tentang bagaimana model klasifikasi bekerja dengan menunjukkan jumlah prediksi yang benar dan salah dibagi berdasarkan kelas. Confusion Matrix membantu dalam menghitung berbagai metrik evaluasi kinerja seperti akurasi, presisi, recall, dan F1-score[16].

   Struktur Confusion Matrix

   Confusion Matrix adalah matriks persegi dengan ukuran $n \times n$, di mana n adalah jumlah kelas dalam dataset. Struktur dasar dari Confusion Matrix untuk model klasifikasi biner adalah sebagai berikut:

	Actually Positive (1)	**Actually Negative (0)**
Predicted Positive (1)	True positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negative (FNs)	True Negatives (TNs)

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

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

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

Dimana,

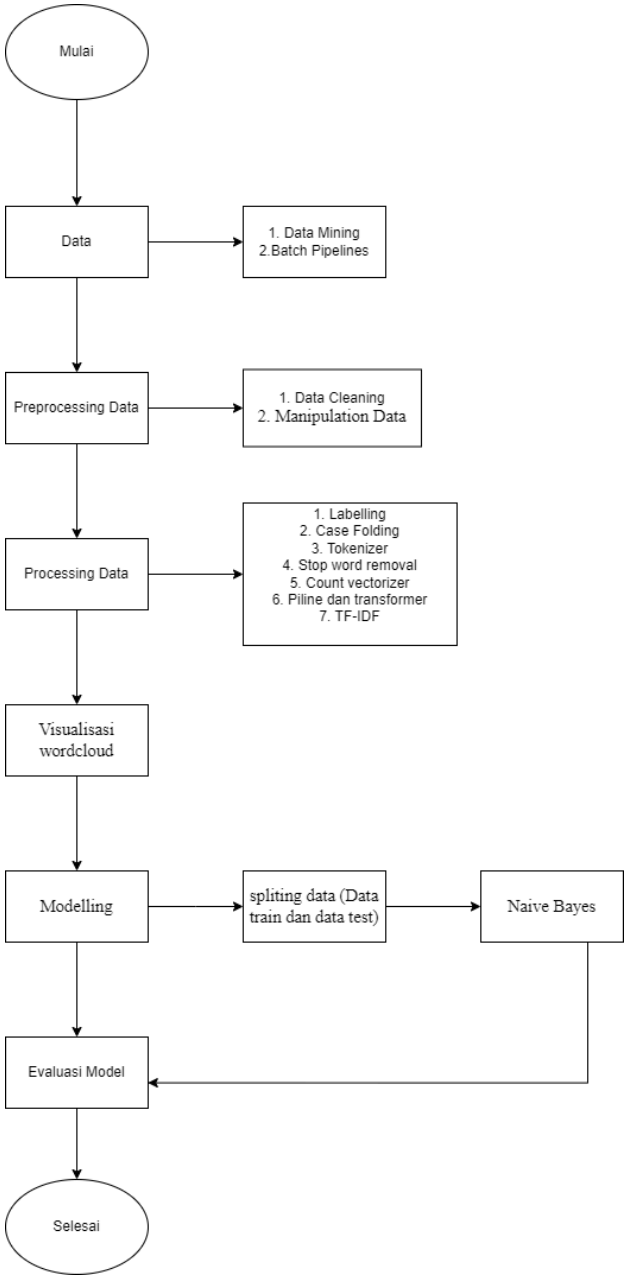
TP : True Positive, yaitu jumlah data positif yang terklasifikasi oleh sistem.

TN : True Negative, yaitu jumlah data bukan positif yang terklasifikasi oleh sistem.

FN : False Negative yaitu jumlah data negatif namun terklasifikasi salah oleh sistem.

FP : False Negative yaitu jumlah data positif namun terklasifikasi salah oleh sistem. [7]

Flowchart



Gambar 2. Alur Penelitian

Hasil dan Diskusi

3.1 Identifikasi Variabel

Data yang digunakan adalah data review pengguna terhadap aplikasi Gojek pada tahun 2022-2024. Variabel terikat pada penelitian ini adalah label pada kolom sentiment berupa positif dan negatif (y), sedangkan variabel bebas pada penelitian ini adalah content atau review pengguna (X_1), score atau rating (X_2).

Tabel 1. Penggunaan variabel pada data review aplikasi Gojek

Variabel	Definisi Operasional	Deskripsi
Y	Labelling	sebuah penanda sentimen yang menggambarkan nilai kata apakah bersifat positif atau negatif
X_1	Content / Review	Berisi ulasan pengguna aplikasi Go-Jek.
X_2	Rating	Sebuah representasi nilai rasio tingkat kepuasan pengguna aplikasi Go-jek

3.2 Model Naive Bayes

Tabel 2. Nilai Akurasi Model

	1	0
1	150919	16469
0	13793	84724

Diterbitkan oleh: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera , Lampung Selatan, Indonesia

Tabel confusion matriks diatas merupakan akurasi dari pemodelan Naive Bayes yang menentukan kelas dari setiap ulasan yang ada, dimana tabel tersebut merepresentasikan apakah setiap kata yang ada di ulasan tersebut mengandung sentimen positif atau negatif dari pengguna aplikasi. Untuk menghitung rumus akurasi dari confusion matriks diatas adalah sebagai berikut.

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

$$Akurasi = \frac{150919 + 16469}{150919 + 13793 + 16469 + 84724} \times 100\%$$

Akurasi = 0.8861924371486057

dapat diartikan sebanyak 88% ulasan pengguna terklasifikasikan dengan benar ke dalam label positif dan negatif.

3.3 Evaluasi Model K-fold Cross Validation

Tabel 3. Nilai Akurasi model k-fold cross validation dengan k=6

	1	0
1	150836	16552
0	13612	84905

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

$$Akurasi = \frac{150836 + 16552}{150836 + 13612 + 16552 + 84905} \times 100\%$$

Akurasi = 0.8868923515871916

Hasil evaluasi penggunaan model k-fold cross validation dengan k =6 diatas dapat diartikan bahwa model Naive Bayes dapat menjelaskan klasifikasi dari ulasan pengguna ke dalam setiap kelas sebesar 88%.

3.4 Visualisasi

a. Word Cloud Positif

Tampilan Word Cloud atau frekuensi kata-kata dari label positif, kata-kata yang paling sering muncul adalah “sangat”, “membantu”, “sekali” seperti pada Gambar 1**.**

Gambar 1. Word Cloud Positif

b. Word Cloud Negatif

Tampilan Word Cloud atau frekuensi kata-kata dari label negatif, kata-kata yang paling sering muncul adalah “tidak”, “bisa” seperti pada **Gambar 2.**

Gambar 2. Word Cloud Negatif

Kesimpulan / Kesimpulan

Berdasarkan hasil penelitian dan pembahasan pada data review pengguna terhadap aplikasi Gojek. Diperoleh kesimpulan bahwa pada pemodelan dengan *Naive Bayes* diperoleh nilai akurasi sebesar 0.8861924371486057, artinya sebanyak 88% ulasan pengguna terklasifikasikan dengan benar ke dalam

label positif dan negatif. Kemudian, pada pemodelan K-fold Cross Validation dengan k=6 diperoleh nilai akurasi sebesar 0.8868923515871916, artinya model Naive Bayes dapat menjelaskan klasifikasi dari ulasan pengguna ke dalam setiap kelas sebesar 88%.

Lampiran

Code

Referensi

[1] H. Zhang, "The Optimality of Naive Bayes," *Journal Faculty of Computer Science,* pp. 1-6, 2004.

[2] Rish, "Anempirical study of the naive Bayes classifier," *Journal T.J. Watson Research Center,* pp. 41-46.

[3] Amazon, "Amazon Web Service," 2023. [Online]. Available: https://aws.amazon.com/id/what-is/apache-spark/. [Diakses 22 Mei 2024].

[4] R. Setiawan, "Dicoding," 30 Oktober 2021. [Online]. Available: https://www.dicoding.com/blog/apa-itu-data-mining/. [Diakses 22 Mei 2024].

[5] Revoupedia, PT Revolusi Cita Edukasi, 2024. [Online]. Available: https://revou.co/kosakata/data-cleaning. [Diakses 22 Mei 2024].

[6] T. Kinasih, "Kuncie," PT. Kuncie Pintar Nusantara, Maret 2023. [Online]. Available: https://www.kuncie.com/posts/apa-itu-data-manipulation/#:~:text=Data%20Manipulation%20atau%20manipulasi%20data,%2C%20menghapus%2C%20dan%20mengubah%20database. [Diakses 22 Mei 2024].

[7] R. Tineges, "DQLab," 17 Juni 2021. [Online]. Available: https://dqlab.id/tahapan-text-preprocessing-dalam-teknik-pengolahan-data. [Diakses 22 Mei 2024].

[8] Y. M. Raditya, "Pembentukan Daftar Stopword Menggunakan Term Based Random Sampling Pada Analisis Sentimen Dengan Metode Naïve Bayes (Studi Kasus: Kuliah Daring Di Masa Pandemi)," *Jurnal Teknologi Informasi dan Ilmu Komputer,* pp. 663-882, 2022.

[9] H. H. J. S. Anasthasya Averina, "Analisis Sentimen Multi-Kelas Untuk Film Berbasis Teks Ulasan Menggunakan Model Regresi Logistik," *TEKNIKA,* pp. 123-128, 2022.

[10] M. A. I. R. H. F. F. S. D. Okta Ihza Gifari, "Analisis Sentimen ReviewFilm Menggunakan TF-IDF dan Support Vector Machine," *JIFOTECH(JOURNAL OF INFORMATION TECHNOLOGY),* pp. 36-41, 2022.

[11] Y. R. Tanjung, "nlimit," Desember 15 2023. [Online]. Available: https://nolimit.id/blog/word-cloud-pengertian-fungsi-dan-cara-membuat/#:~:text=Word%20cloud%20merupakan%20visualisasi%20dari,disebutkan%20para%20pengguna%20media%20sosial. [Diakses 22 Mei 2024].

[12] R. Tineges, "DQLab," 23 Mei 2022. [Online]. Available: https://dqlab.id/mengenal-naive-bayes-sebagai-salah-satu-algoritma-data-science. [Diakses 22 Mei 2024].

Diterbitkan oleh: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera , Lampung Selatan, Indonesia

[13] I. C. B. R. Reza Aprilliana Fauzi, "Pemanfaatan Spark untuk Analisis Sentimen Mengenai Netralitas Berita dalam Membahas Pemilu Presiden 2019 Menggunakan Metode Naïve Bayes Classifier," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* pp. 1070-1077, 2021.

[14] E. H. N. I. Alfio Kusuma, "AnalisisSentimen Pada Ulasan Aplikasi Indodax di Google Play StoreMenggunakan Metode Support Vector Machine," *SeminarNasionalMahasiswaIlmuKomputerdan Aplikasinya(SENAMIKA),* pp. 563-573, 2022.

[15] V. C. M. T. S. Muhammad Farras, "Aplikasi Analisis Sentimen Komentar Pengguna Genshin Impact Di Play Store," *Jurnal Ilmu Komputer dan Sistem Informasi,* pp. 1-6.

[16] N. H. R. M. H. H. N. S. R. E. MuhammadNurAkbar, " Sentiment Analysis Terhadap Review Aplikasi Maxim di Google Play Store Menggunakan Support Vector Machine (SVM)," *Journal of Artificial Inteligence & Data Science,* pp. 1-8, 2022.

[17] P. S. B. A. K. Primandani Arsi, "Analisis Sentimen Game Genshin Impact pada Play Store Menggunakan Naïve Bayes Clasifier," *JURNAL ILMIAH TEKNIK MESIN, ELEKTRO DAN KOMPUTER,* pp. 162-169, 2023.