



Analisis Sentimen Access by Bus Kota se-Indonesia Menggunakan Metode K-Nearest Neighbors

M. Andrik Muqorrobin P^{1*}, Zaehol Fatah²

^{1,2}Fakultas Sains dan Teknologi, Universitas Ibrahimy Situbondo, Jl. KHR. Syamsul Arifin No.1-2, Situbondo 68374, Indonesia

¹wmmamuqorrobinp@gmail.com^{1*}; ²zaeholfatah@gmail.com

Korespondensi penulis: wmmamuqorrobinp@gmail.com

Abstract; Data mining is the process of collecting and processing data to extract important information. The K-Nearest Neighbor data mining method can analyze the Redbus application. Access by Bus Kota is one of the official applications for purchasing city bus tickets in Indonesia. The problem that arises after the update of the Access by Bus Kota application is the increase in one-star reviews stating that the latest version does not match the previous version. Data Mining is used to analyze the sentiment of Access by Bus Kota throughout Indonesia using the K-Nearest Neighbors method. The data used is data obtained from redBus application user reviews for one month starting from September 20, 2024 to October 20, 2024 with a total of 1291 reviews. Sentiment analysis in this study uses the K-Nearest Neighbors method through the Python programming language. The results showed that the best performance in experiments with the division of training data and test data, as well as varying k values was obtained in experiments with a division of 90% training data, 10% test data and using a value of k = 5 with accuracy, precision, and recall values of 90.23%; and a recall value of 72.38%, respectively. Sentiment classification with the best model using parameter k = 3 produces 79.26% positive sentiment, 17.25% neutral sentiment, and 3.49% negative sentiment.

Keywords: Analysis, K-nearest Neighbors, Sentiment

Abstrak; Data mining atau penambangan data merupakan proses pengumpulan dan pengolahan data untuk mengekstrak informasi penting. Metode data mining K-Nearest Neighbor dapat menganalisis pada aplikasi Redbus. RedBus merupakan salah satu aplikasi resmi pembelian tiket bus kota di Indonesia. Permasalahan yang muncul setelah pembaruan aplikasi RedBus adalah bertambahnya ulasan bintang satu yang menyatakan bahwa versi terbaru tidak sesuai dengan versi sebelumnya. Data Mining yang digunakan untuk menganalisis sentimen Access by Bus Kota di seluruh Indonesia menggunakan metode K-Nearest Neighbors. Data yang digunakan adalah data yang diperoleh dari ulasan pengguna aplikasi redBus selama satu bulan terhitung dari tanggal 20 September 2024 sampai dengan 20 Oktober 2024 dengan total 1291 ulasan. Analisis sentimen pada penelitian ini menggunakan metode K-Nearest Neighbors melalui bahasa pemrograman Python. Hasil penelitian menunjukkan bahwa kinerja terbaik pada percobaan dengan pembagian data latih dan data uji, serta nilai k yang bervariasi diperoleh pada percobaan dengan pembagian 90% data latih, 10% data uji dan menggunakan nilai k = 5 dengan nilai akurasi, presisi, dan recall masing-masing sebesar 90,23%; dan nilai recall sebesar 72,38%. Klasifikasi sentimen dengan model terbaik menggunakan parameter k = 3 menghasilkan 79,26% sentimen positif, 17,25% sentimen netral, dan 3,49% sentimen negatif.

Kata kunci: Analisis, K-nearest Neighbors, Sentimen.

1. Pendahuluan

Transportasi daring merupakan salah satu bentuk transportasi dan lalu lintas jalan raya yang memanfaatkan dan mengikuti kemajuan ilmu pengetahuan (teknologi) melalui aplikasi mobile dan pembelian serta pembayaran secara daring (Oladimeji et al., 2023). Penemuan ini diciptakan untuk mengatasi sejumlah masalah dalam transportasi umum tradisional, termasuk kemudahan pemesanan, ketersediaan, dan digitalisasi pembayaran (Mogaji & Nguyen, 2024). Kemajuan teknologi yang pesat memiliki dampak besar pada masyarakat (Volti & Croissant, 2024). Meningkatnya popularitas transportasi online melalui internet merupakan salah satu konsekuensi dari kemajuan teknologi (Mouratidis et al., 2021). Di Asia Tenggara, Indonesia mengalami peningkatan pangsa pasar transportasi online terbesar (Liza et al., 2024). Menurut survei Asosiasi Penyelenggara Jasa Internet Indonesia, 93,9%

masyarakat menggunakan smartphone untuk mengakses internet, dan reservasi transportasi online menduduki peringkat ke-16 dari 22 alasan masyarakat menggunakan internet. (Prananda et al., 2020).

Aplikasi seluler telah muncul sebagai salah satu instrumen terpenting dalam kehidupan sehari-hari di era digital yang terus berkembang (Mois & Rogers, 2024). Aplikasi ini berdampak pada sejumlah hal, termasuk layanan publik seperti transportasi bus kota (Javid et al., 2021). "Access by Bus Kota," juga dikenal sebagai redBus, adalah salah satu aplikasi resmi untuk membeli tiket bus kota dan memperoleh informasi perjalanan. Aplikasi ini merupakan salah satu aplikasi yang perlu dipertimbangkan dalam hal ini. Redbus berdiri pada tahun 2006 dan telah menghubungkan berbagai kota di Indonesia. Masalah yang muncul pasca pembaruan aplikasi redBus adalah meningkatnya ulasan bintang satu yang menyatakan bahwa versi terbaru tidak sesuai dengan versi sebelumnya. Sebelum adanya perubahan, banyak pengguna yang menyatakan lebih memilih menggunakan program versi sebelumnya.

Banyaknya ulasan dan tanggapan pengguna di aplikasi redBus di Google Playstore menunjukkan adanya kecenderungan sentimen terhadap layanan pemesanan bus kota tersebut. Teknik perolehan, pemrosesan otomatis, dan pemahaman data dalam bentuk teks tak terstruktur untuk mengekstrak informasi sentimen dari frasa opini atau pendapat dikenal dengan analisis sentimen (Wankhade et al., 2022). Tujuan dasar dari analisis sentimen adalah untuk mengklasifikasikan polaritas teks, yaitu apakah pendapat yang diungkapkan bersifat netral, negatif, atau positif (Abd et al., 2021). Salah satu dari banyak teknik yang tersedia untuk analisis sentimen adalah K-Nearest Neighbor (K-NN) (Nuraini et al., 2024). K-NN adalah metode yang mudah diimplementasikan dan sederhana yang menggunakan data berlabel untuk membantu mengelompokkan data ke dalam kelas yang paling tepat (Alnuaimi & Albaldawi, 2024). Manfaatnya meliputi kemampuan untuk mengklasifikasikan data dengan data pelatihan dan pengujian, kemudahan dalam menafsirkan hasil, dan keakuratan prediksi dengan mengurutkan nilai k terdekat terlebih dahulu.

Mengenai analisis sentimen menggunakan metode K-Nearest Neighbor (K-NN), ada beberapa penelitian yang telah dilakukan sebelumnya. Iwandini et al., (2023) meneliti tentang analisis sentimen pengguna transportasi Jakarta terhadap Transjakarta dengan menggunakan metode Naives Bayes dan K-Nearest Neighbor. Metode yang digunakan untuk klasifikasi sentimen ini adalah perbandingan metode Naive Bayes Classifier dan juga K-Nearest Neighbor Classifier dengan pembobotan TF-IDF. Input dari sistem ini berupa data tweet tentang Transjakarta, dan output dari sistem ini berupa visualisasi data sentimen positif dan negatif menggunakan Streamlit yang merupakan library dari python. Berdasarkan pengujian, akurasi pendekatan Naive Bayes untuk analisis sentimen data Twitter terkait penggunaan transportasi Transjakarta adalah sebesar 61,1%, dan akurasi metode K-Nearest Neighbor adalah sebesar 75,7%. Untuk kedua metode yang digunakan dalam menentukan tingkat akurasi, dapat diketahui bahwa metode K-nearest-neighbor menghasilkan akurasi yang lebih tinggi.

Penelitian selanjutnya dilakukan oleh Hakim & Sugiyono, (2024) tentang analisis sentimen kereta cepat Jakarta Bandung menggunakan algoritma Naïve Bayes dan K-Nearest Neighbor. Hasil penelitian mereka menemukan bahwa mayoritas masyarakat Indonesia memiliki respon negatif terhadap kereta cepat Jakarta-Bandung dan berdasarkan klasifikasi model algoritma K-Nearest Neighbor, dengan menggunakan rasio split data sebesar 0,8:0,2 dengan nilai k=3 terhadap dataset kereta cepat Jakarta-Bandung diperoleh nilai akurasi sebesar 99,76%. Sehingga dapat dikatakan bahwa algoritma KKNN dapat mengklasifikasikan data dengan baik dan benar.

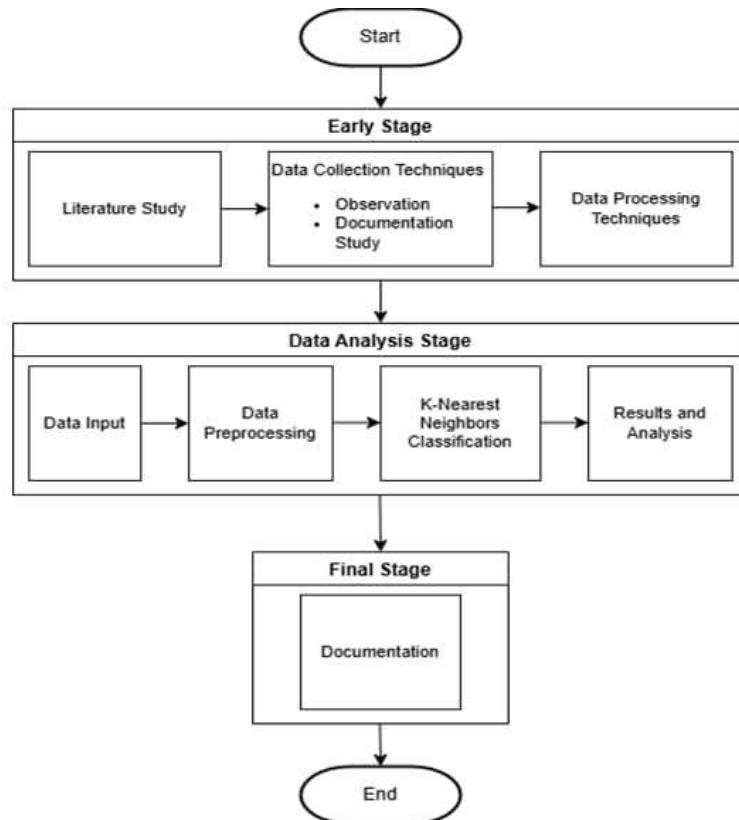
Terakhir, penelitian yang dilakukan oleh Irawan et al., (2022) tentang analisis sentimen pengguna Gojek menggunakan metode K-Nearest Neighbour. Hasil penelitian menunjukkan bahwa hasil klasifikasi menggunakan metode K-Nearest Neighbor mampu mengklasifikasikan respon dari pengguna Twitter dan dapat digunakan oleh perusahaan Gojek sebagai bahan evaluasi dan penilaian layanan Gojek. Hasil pengujian metode K-Nearest Neighbor menggunakan matriks konfusi dengan 1409 data diperoleh tingkat akurasi sebesar 79,43% dengan nilai k = 15.

Mengacu pada beberapa fenomena dan penelitian sebelumnya, tujuan dari penelitian ini adalah menganalisis sentimen Akses Bus Kota di seluruh Indonesia menggunakan metode K-Nearest Neighbors.

2. Metode Penelitian

Tahapan penelitian ini digambarkan melalui diagram alir metodologi penelitian pada Gambar 1. Tahap awal dalam penelitian adalah melakukan studi pustaka untuk mencari landasan teori dan referensi

tertulis yang mendasari penelitian agar dapat menyelesaikan permasalahan dan dijadikan acuan yang kuat dalam melakukan analisis. Pengolahan data diawali dengan (a) Tahap praproses data; (b) Pembagian data latih dan data uji; (c) Pemodelan klasifikasi data dengan K-NN; dan (d) Analisis sentimen hasil klasifikasi. Tahapan praproses data review terdiri dari case folding, tokenizing, cleansing, stopword removal, stemming, pelabelan data review, dan diakhiri dengan proses pembobotan TF-IDF. Setelah data dipraproses, dilakukan proses distribusi data latih dan data uji dengan besaran yang bervariasi, kemudian dilakukan pemodelan data dengan KNN menggunakan beberapa nilai parameter k. Data yang digunakan adalah data yang diperoleh dari ulasan pengguna aplikasi redBus selama satu bulan terhitung dari tanggal 20 September 2024 sampai dengan 20 Oktober 2024 dengan total 1291 ulasan. Data ulasan pengguna aplikasi redBus diperoleh dari Google Play Store yang dapat diakses melalui tautan <https://play.google.com/store/apps/details?id=in.redbus.android&hl=id&pli=1> dengan menggunakan teknik scraping melalui bahasa pemrograman Python. Proses scraping data ulasan menggunakan Package `google_play_scraper` yang tersedia dalam bahasa python.



Gambar 1. Diagram Alir Penelitian

a. Analisis Sentimen

Analisis sentimen merupakan bagian dari pemrosesan bahasa alami (Natural Language Processing, NLP) yang berfokus pada pengenalan dan pengelompokan opini atau perasaan yang terdapat dalam suatu teks. Analisis ini biasanya digunakan untuk mengidentifikasi apakah sebuah teks memiliki sentimen positif, negatif, atau netral. Dalam penelitian ini, analisis sentimen dilakukan untuk menggali pandangan masyarakat mengenai layanan transportasi bus kota di seluruh Indonesia.

b. Algoritma K-Nearest Neighbor

K-Nearest Neighbor adalah salah satu metode supervised learning yang melibatkan proses pembelajaran dari data training. Algoritma K-NN sangat umum digunakan dalam melakukan

klasifikasi objek karena pengaplikasiannya sangat sederhana, efektif dan efisien.(Sutrisno & Amini, 2023)

a) Mengukur jarak antara data yang diklasifikasikan dengan data yang ada dalam dataset. Metode perhitungan jarak yang umum dilakukan adalah:

- Euclidean Distance

$$d(p,q)=\sqrt{\sum_{i=1}^n (q_i-p_i)^2}$$

- Manhattan Distance

$$d(p,q)=\sum_{i=1}^n |q_i-p_i|$$

b) Menentukan tetangga terdekat berdasarkan k data yang ditemukan.

c) Data baru yang diklasifikasikan ke dalam kelas yang sering muncul di antara k tetangga tersebut.

c. Pembobotan TF IDF

Model vektor yang telah ditentukan akan dijalankan pembobotan lewat TF – IDF ini agar jumlah kemunculan suatu kata dalam kumpulan dataset bisa ditemukan dan diketahui. Banyaknya kemunculan suatu kata bisa diketahui lewat Term Frequency (TF) sementara hasil perkalian TF dan IDF untuk menemukan jumlah kemunculan suatu kata bisa dilihat da diketahui lewat Inverse Document Frequency (Cahyani N. Dkk 2022).

d. Klasifikasi K-Nearest Neighbor

Klasifikasi merupakan teknik untuk mengklasifikasikan suatu data tertentu ke dalam kelas yang sebelumnya telah ditentukan kelasnya pada pola atau atribut yang terdapat pada data tersebut. Klasifikasi ini menjadi lebih popular dikarenakan dapat menangani rentang data yang lebih luas dibandingkan dengan regresi. Tujuan utama dari klasifikasi ini sendiri yakni untuk memprediksi kelas atau label dari data yang belum dikenal berdasarkan data yang telah ada dari data yang sebelumnya telah diketahui.(Homaidi & Fatah, 2024)

e. Rapid Miner

Rapid Miner adalah suatu Analisis teks yang bekerja untuk memproses data mining, dengan melibatkan penggalian pola dari kumpulan data besar dan menggabungkannya dengan Teknik statistik, kecerdasan buatan, dan basis data. Tujuan dari analisis teks ini adalah untuk menghasilkan informasi yang akurat dan bermakna dari teks yang sangat besar atau kompleks.(Natasya & Fatah, 2024)

3. Hasil dan Pembahasan

Data review pada Tabel 1. yang diperoleh dari hasil scraping tidak dapat langsung digunakan dalam proses mining dengan algoritma K-Nearest Neighbors (KNN) Data Mining. Data review perlu dilakukan preprocessing untuk menghilangkan data noise, missing value, atau data review yang tidak memiliki informasi yang berguna, tidak memiliki makna, dan data yang dapat mengganggu proses mining.

1. Mengumpulkan Data

Mengumpulkan data merupakan kegiatan mencari data informasi yang diperoleh guna mendukung tujuan suatu penelitian. Proses ini melibatkan pengumpulan berbagai jenis data dari sumber tertentu menggunakan metode yang sesuai kualitatif maupun kuantitatif.

A	B	C	D
Tanggal	Teks	Sentimen	Lokasi
1 2024-03-11	Pengalaman naik bus di Yogyakarta men...	Positif	Yogyakarta
2 2023-12-01	Akses bus di Makassar sangat memudah...	Positif	Makassar
3 2023-12-10	Layanan bus di Palembang sangat nyam...	Positif	Palembang
4 2024-08-31	Akses bus di Yogyakarta tidak memadai...	Negatif	Yogyakarta
5 2023-12-13	Bus di Medan sering tidak sesuai jadwal...	Negatif	Medan
6 2024-02-12	Jadwal bus di Bandung sangat teratur...	Positif	Bandung
7 2023-12-08	Bus di Balikpapan sering terlambat dan p...	Negatif	Balikpapan
8 2024-04-27	Akses bus di Semarang sangat memudah...	Positif	Semarang
9 2024-09-25	Akses bus di Balikpapan sangat memudah...	Positif	Balikpapan
10 2023-12-30	Akses bus di Balikpapan sangat memudah...	Positif	Balikpapan
11 2024-08-05	Pengalaman naik bus di Bandung menge...	Negatif	Bandung
12 2023-11-09	Layanan bus di Medan buruk dan tidak te...	Negatif	Medan
13 2023-11-08	Akses bus di Palembang tidak memadai...	Negatif	Palembang
14 2024-10-31	Bus di Medan bersih dan nyaman...	Positif	Medan
15 2024-10-01	Layanan bus di Yogyakarta buruk dan tida...	Negatif	Yogyakarta
16 2024-09-08	Bus di Balikpapan sering terlambat dan p...	Negatif	Balikpapan
17 2024-08-07	Bus di Yogyakarta bersih dan nyaman...	Positif	Yogyakarta
18 2024-10-14	Jadwal bus di Balikpapan sangat teratur...	Positif	Balikpapan
19 2024-02-28	Bus di Makassar bersih dan nyaman...	Positif	Makassar
20 2024-03-01

Gambar 2. Ulasan Pengguna Aplikasi redBus

2. Divisi Data Pelatihan dan Data Uji

Tahapan praproses data tinjauan terdiri dari pelipatan kasus, tokenisasi, pembersihan, penghapusan stopword, stemming, pelabelan data tinjauan, dan diakhiri dengan proses pembobotan TF-IDF.

a. Case Holding Stage

A	B	C
Ulasan	Case Folding	
1 bus di jember sering TERLAMBAT dan PENUH.	bus di jember sering terlambat dan penuh.	
2 bus di surbaya sering terlambat dan PENUH.	bus di surabaya sering terlambat dan penuh.	
3 bus di bandung sering trimbat dan penuh.	bus di bandung sering terlambat dan penuh.	
4 bus di balikpapi sering triambat dan penuh.	bus di balikpapan sering terlambat dan penuh.	
5 akses bus di jkt tidak memadai	akses bus di jakarta tidak memadai	
6 bus di makassar sering tdk sesuai jadwal	bus di makassar sering tidak sesuai jadwal	
7 akses bus di sibond singat memudahkan perjman.	akses bus di stibondo sangat memudahkan perjalanan.	
8 layanan bus di bandung BURUK dan TIDAK tepat waktu.	layanan bus di bandung buruk dan tidak tepat waktu.	
9 bus di semarang sering TIDAK sesuai jadwal.	bus di semarang sering tidak sesuai jadwal.	
10 akses bus di palembang singat memudahkan perjman.	akses bus di palembang sangat memudahkan perjalanan	
11 jdw bus di BANTEN sangat teratur.	jadwal bus di banten sangat teratur.	
12 layann bus di jogya sangat nyaman dan tpat wktu.	layanan bus di yogyakarta sangat nyaman dan tepat wakt.	
13 bus di BONDOWOSO sering terlambat dan penuh.	bus di bondowoso sering terlambat dan penuh.	
14 bus di srby sering TERLAMBAT dan penuh.	bus di surabaya sering terlambat dan penuh.	
15 bus di bandung sering terlambat dan penuh.	bus di bandung sering terlambat dan penuh.	
16 bus di balikpapi sering terlambat dan penuh.	bus di balikpapan sering terlambat dan penuh.	
17 akses bus di jkt tidak memadai	akses bus di jakarta tidak memadai	
18 bus di makassar sering tidak sesuai jadwal	bus di makassar sering tidak sesuai jadwal	
19 akses bus di jogja sangat memudahkan perjalanan.	akses bus di yogyakarta sangat memudahkan perjalanan.	
20		

Gambar 3. Data Hasil Case Folding

b. Tokenizing Stage

Tahap tokenisasi adalah proses memecah kalimat atau string menjadi token dalam bentuk karakter atau fragmen kata (Yang, 2024). Misalnya, kalimat "titik jemput akurat" akan dipecah menjadi fragmen kata "titik", "jemput", dan "akurat". Proses tokenisasi dalam ulasan ini

menggunakan pustaka nltk yang tersedia dalam bahasa python. Berikut ini adalah hasil data yang menjalani proses tokenisasi pada Tabel 3.

c. Cleansing Stage

Tahap Cleansing terbagi menjadi dua tahap, tahap pertama adalah proses mengubah kata-kata slang atau kata-kata tidak baku pada data review, proses normalisasi mengacu pada kamus kata normalisasi yang digunakan untuk mengubah data-data tidak baku menjadi kata-kata baku atau normal. Seperti kata “7an” dinormalisasi menjadi kata “tujuan”; kata “abis” menjadi “habis”; kata “adlah” menjadi “adalah”, dan seterusnya. Tahap selanjutnya adalah tahap filtering yaitu proses menghilangkan karakter-karakter pada suatu string yang bukan merupakan bagian dari alfabet dan tidak memiliki makna tertentu sehingga dapat mengurangi karakter-karakter yang tidak diperlukan, seperti angka; hashtag; emoji; link; tanda baca, dan sebagainya. Berikut ini adalah hasil akhir dari proses cleaning data review pada Tabel 3.

A	B
1 Tokenisasi	Cleansing
2 ["bus", "di", "Tokenisasi", "j", "terlambat", "dan", "penuh"]	bus di jember sering terlambat dan penuh
3 ["bus", "di", "surabaya", "sering", "terlambat", "dan", "penuh"]	bus di surabaya sering terlambat dan penuh
4 ["bus", "di", "bandung", "sering", "terlambat", "dan", "penuh"]	bus di bandung sering terlambat dan penuh
5 ["bus", "di", "balikpapan", "sering", "terlambat", "dan", "penuh"]	bus di balikpapan sering terlambat dan penuh
6 ["akses", "bus", "di", "jakarta", "tidak", "memadai"]	akses bus di jakarta tidak memadai
7 ["bus", "di", "makassar", "sering", "tidak", "sesuai", "jadwal"]	bus di makassar sering tidak sesuai jadwal
8 ["akses", "bus", "di", "situbondo", "sangat", "memudahkan", "perjalanan"]	akses bus di situbondo sangat memudahkan perjalanan
9 ["layanan", "bus", "di", "bandung", "buruk", "dan", "tidak", "tepat", "waktu"]	layanan bus di bandung buruk dan tidak tepat waktu
10 ["bus", "di", "semarang", "sering", "tidak", "sesuai", "jadwal"]	bus di semarang sering tidak sesuai jadwal
11 ["akses", "bus", "di", "palembang", "sangat", "memudahkan", "perjalanan"]	akses bus di palembang sangat memudahkan perjalanan
12 ["jadwal", "bus", "di", "banten", "sangat", "teratur"]	jadwal bus di banten sangat teratur
13 ["layanan", "bus", "di", "yogyakarta", "sangat", "nyaman", "dan", "tepat", "waktu"]	layanan bus di yogyakarta sangat nyaman dan tepat waktu
14 ["bus", "di", "bondowoso", "sering", "terlambat", "dan", "penuh"]	bus di bondowoso sering terlambat dan penuh
15 ["bus", "di", "surabaya", "sering", "terlambat", "dan", "penuh"]	bus di surabaya sering terlambat dan penuh
16 ["bus", "di", "bandung", "sering", "terlambat", "dan", "penuh"]	bus di bandung sering terlambat dan penuh
17 ["bus", "di", "balikpapan", "sering", "terlambat", "dan", "penuh"]	bus di balikpapan sering terlambat dan penuh
18 ["akses", "bus", "di", "jakarta", "tidak", "memadai"]	akses bus di jakarta tidak memadai
19 ["bus", "di", "makassar", "sering", "tidak", "sesuai", "jadwal"]	bus di makassar sering tidak sesuai jadwal
20 ["akses", "bus", "di", "yogyakarta", "sangat", "memudahkan", "perjalanan"]	akses bus di yogyakarta sangat memudahkan perjalanan

Gambar 4. Data Hasil Tokenizing & Cleansing

d. Stopword Removal Stage

Tahap stopword removal merupakan tahap untuk membuang kata-kata umum yang tidak memiliki makna tertentu, seperti kata hubung atau kata sambung dan kata keterangan atau adverbial yang tidak memberikan pengaruh pada tahap analisis mining. Seperti kata “saya”, “dia”, “di bawah”, “dengan”, “nya”, “itu”, “ini”, dan kata-kata lainnya. Tahapan penghilangan stopword menggunakan kamus stopword yang mengacu pada stoplist pada penelitian Irawan et al., (2022) dan kata-kata lainnya merupakan hasil pengumpulan yang dilakukan oleh peneliti. Berikut ini adalah hasil tahapan penghilangan stopword dengan menyisihkan data akhir sebanyak 4269 data mentah seperti pada Tabel 4.

e. Stemming Stage

Proses stemming dalam penelitian ini menggunakan pustaka Sastrawi yang tersedia dalam bahasa Python. Proses ini bertujuan untuk mengubah kata ke bentuk dasar sesuai dengan kaidah dalam Kamus Besar Bahasa Indonesia (KBBI), misalnya kata “menunggu” menjadi kata dasar “tunggu”; kata membantu menjadi kata “bantu”; dan seterusnya. Hasil dari tahap stemming tercantum dalam Tabel 4.

A	B	C
Teks	Stopword Removal	Stemming
1 "Pelayanan bus sangat memuaskan, nyaman dan tepat waktu"	Pelayanan bus sangat memuaskan, nyaman tepat waktu	Layanan bus sangat puas, nyaman tepat waktu
2 "Terlalu sering telat, sangat mengecewakan"	Terlalu sering telat, sangat mengecewakan	Terlalu sering telat, sangat kecewa
3 "Harga tiket sangat mahal, tidak sebanding dengan kualitasnya"	Harga tiket sangat mahal, sebanding kualitasnya	Harga tiket sangat mahal, banding kualitas
4 "Bus bersih dan nyaman, sangat direkomendasikan"	Bus bersih nyaman, sangat direkomendasikan	Harga tiket sangat mahal, banding kualitas
5 "Sopir ramah dan sopan, pelayanan sangat baik"	Sopir ramah sopan, pelayanan sangat baik	Sopir ramah sopan, layanan sangat baik
6 "Bus sering telat, sangat tidak nyaman"	Bus sering telat, tidak nyaman	Bus sering telat, tidak nyaman
7 "Pelayanan memuaskan, bus baru dan bersih"	Pelayanan memuaskan, bus baru bersih	Layanan puas, bus baru bersih
8 "Kondisi bus buruk, banyak yang rusak"	Kondisi bus buruk, banyak rusak	Kondisi bus buruk, banyak rusak
9 "Bus sangat nyaman, AC dingin, sopir ramah"	Bus sangat nyaman, AC dingin, sopir ramah	Bus sangat nyaman, AC dingin, sopir ramah
10 "Layanan bus buruk, sering terlambat dan tidak ada informasi"	Layanan bus buruk, sering terlambat tidak ada informasi	Layanan bus buruk, sering terlambat tidak ada info
11 "Bus baru, sangat modern dan efisien"	Bus baru, sangat modern efisien	Bus baru, sangat modern efisien
12 "Terlalu banyak penumpang, sangat sesak"	Banyak penumpang, sangat sesak	Banyak penumpang, sangat sesak
13 "Sangat disarankan untuk naik bus ini jika ingin perjalanan lancar"	Disarankan naik bus ini jika ingin perjalanan lancar	Saran naik bus ini jika ingin perjalanan lancar
14 "Jadwal bus tidak tepat, sering berubah-ubah"	Jadwal bus tidak tepat, sering berubah-ubah	Jadwal bus tidak tepat, sering ubah
15 "Tolong perbaiki kondisi bus, banyak kerusakan di dalamnya"	Perbaiki kondisi bus, banyak kerusakan dalamnya	Perbaiki kondisi bus, banyak rusak dalam
16 "Pelayanan sangat buruk, sopir tidak ramah dan tidak sopan"	Pelayanan buruk, sopir tidak ramah tidak sopan	Layanan buruk, sopir tidak ramah tidak sopan

Gambar 5. Stopword Removal and Stemming Results

f. Labeling Stage

Tahap pelabelan merupakan data teks yang tidak terstruktur dan tidak memiliki kelas/label, sehingga untuk dapat digunakan dalam proses klasifikasi dengan algoritma K-NN diperlukan proses pelabelan data. Proses pelabelan data memerlukan kamus kata sentimen positif dan negatif. Kamus kata negatif dan positif yang digunakan dalam penelitian ini adalah kamus yang dibuat berdasarkan referensi penelitian Jaman et al., (2020) dan dilengkapi dengan kamus yang terdapat pada repositori milik Prakoso, (2017) dan berdasarkan kajian pustaka yang dilakukan oleh peneliti. Tahapan dalam pemberian label pada data teks adalah sebagai berikut Harahap et al., (2024):

- Menentukan istilah-istilah yang merujuk dan mampu mewakili kata positif dan negatif, dalam hal ini adalah adanya kamus kata-kata sentimen positif dan negatif. Kata positif adalah kata-kata yang menggambarkan sanjungan, pujian, atau yang memiliki makna baik dan terpuji, sedangkan kata negatif adalah kata-kata yang mewakili kata-kata yang tidak memiliki atau kurang memiliki makna baik yang biasanya mengarah pada antonim atau negasi dari kata-kata pujian atau positif.
- Menghitung jumlah kata positif dan negatif pada setiap kalimat, dimana apabila jumlah kata positif lebih banyak daripada jumlah kata negatif pada kalimat tersebut, maka kalimat tersebut masuk ke dalam kelas sentimen positif, sebaliknya apabila jumlah kata positif lebih sedikit daripada jumlah kata negatif pada kalimat tersebut, maka kelas kalimat tersebut masuk ke dalam kelas sentimen negatif. Sedangkan apabila jumlah kata positif sama dengan jumlah kata negatif pada suatu kalimat, maka kalimat tersebut masuk ke dalam kelas sentimen netral.

A	B	C
1 Stemming	Count	Sentimen
2 Lokasi	1	Positive
3 Dvr ramah	1	Positive
4 bagus baik	2	Positive
5 Terima kasih driver	3	Positive
6 Cike terima kasih	1	Positive
7 Membantu ramah sopimya	2	Positive
8 Tolong baik tunggu	-1	Negative
9 Puas	0	Neutral
10 baik	2	Positive
11 Terima kasih santun baik	3	Positive
12 Bagus	2	Positive

Gambar 6. Labeling Result Data

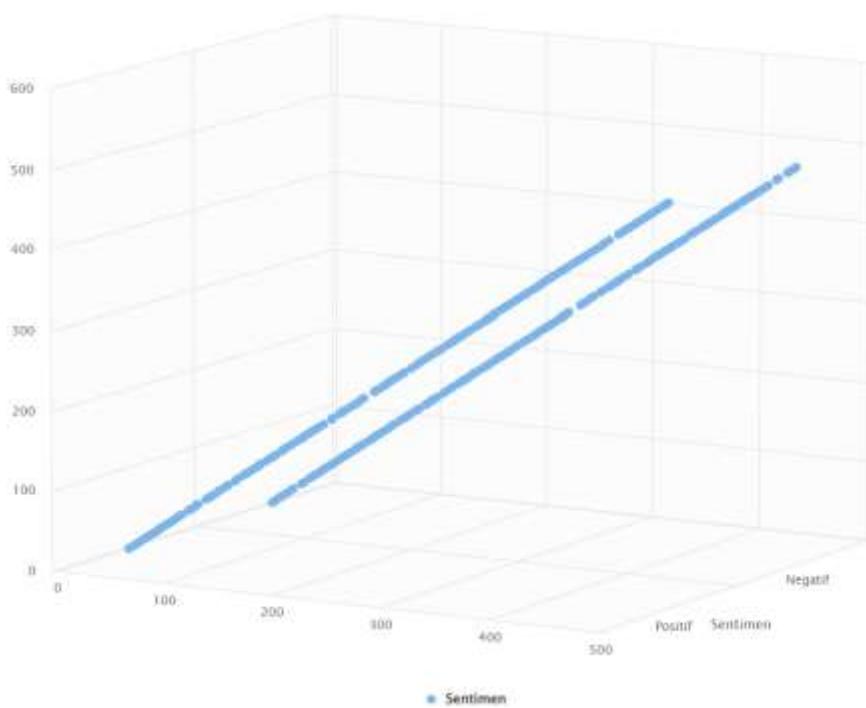
g. TF-IDF Weighting Stage

Analisis frekuensi kata TF-IDF (*Term Frequency-Inverse Document Frequency*) digunakan untuk mengimplementasikan pembobotan kata yang bekerja dengan menghitung frekuensi setiap kata dan melaporkan frekuensi kata-kata yang paling penting, serta frekuensi kemunculan kata dalam dokumen. Perhitungan TF-IDF dapat dilihat pada persamaan (1).

$$W_{dt} = TF_{dt} \times \log \frac{d}{df}$$

3. Best Model Evaluation

Pemodelan klasifikasi sentimen dengan algoritma KNN menggunakan library scikit learn dengan modul klasifikasi K-Nearest Neighbors (Patil et al., 2024). Pemodelan pada penelitian ini menggunakan nilai parameter k atau jumlah tetangga terdekat yang bervariasi yaitu k = 3, k = 5, dan k = 7. Dalam mengevaluasi model yang dibuat, dilakukan perhitungan kinerja model menggunakan modul metrik pada python.



Gambar 7. Hasil Uji Evaluasi K-NN dengan Berbagai K pada Data Resampling

Berdasarkan nilai evaluasi pada setiap skenario percobaan dengan data resampling SMOTE, model dengan kinerja terbaik menghasilkan nilai akurasi sebesar 88,78%; presisi sebesar 88,78%; dan recall sebesar 88,89% pada percobaan dengan nilai $k = 5$ dan 70% data latih serta 30% data uji (P-3). Sebaliknya, model dengan kinerja terendah diperoleh pada percobaan dengan nilai $k = 3$ dan data uji serta data latih masing-masing sebesar 75% dan 25% (P-4) dengan nilai akurasi; presisi; dan nilai recall masing-masing sebesar 81,63%; 81,63%; dan 81,75%. Berdasarkan 2 pemodelan algoritma KNN tanpa dan menggunakan teknik resampling SMOTE dapat disimpulkan bahwa pemodelan dengan nilai evaluasi terbaik diperoleh pada pemodelan tanpa penerapan SMOTE pada proses rekonstruksi data latih dan data uji, dimana penggunaan SMOTE tidak berpengaruh terhadap peningkatan nilai evaluasi pada klasifikasi data review pengguna redBus sebagai representasi pengguna Access by Bus Kota di seluruh Indonesia.

Selain itu dilakukan evaluasi model terhadap model yang menghasilkan nilai kinerja terbaik dalam hal ini adalah model dengan parameter $k = 3$ dan proporsi data latih sebesar 90% serta data uji sebesar 10%. Berikut ini merupakan perbedaan perbedaan kelas berdasarkan kelas pada data uji aktual dan kelas pada data uji berdasarkan hasil prediksi dengan model KNN.

A	B	C	D	E
Kelas Data	Sentimen			Jumlah Uji Data
2	Positif	Netral	Negatif	
3 Kelompok Aktual	767.000	135.000	72.000	974.000
4 Kelompok Prediksi	759.000	163.000	32.000	954.000

Gambar 8. Hasil Uji Model Terbaik

Jika dilihat dari Tabel 6 dapat disimpulkan bahwa model cenderung menghasilkan kelas positif dan netral. Hal ini dapat dilihat dari hasil prediksi yang menghasilkan sentimen netral lebih banyak daripada kelas aktual, begitu pula sentimen positif pada hasil prediksi. Di sisi lain, sentimen negatif justru mengalami penurunan dari data aktual berdasarkan hasil prediksi dari 72 review negatif menjadi 32 review negatif saat diprediksi menggunakan model KNN. Dengan demikian dapat disimpulkan bahwa pada percobaan dengan parameter $k = 3$ terdapat 79,26% sentimen positif, 17,25% sentimen netral, dan 3,49% sentimen negatif.

4. Sentiment Analysis Reviews

Sentimen dapat dianalisis dengan memvisualisasikan menggunakan pustaka wordcloud dan matplotlib yang tersedia dalam bahasa Python. Berdasarkan proses visualisasi, diperoleh wordcloud sentimen netral dari ulasan pengguna aplikasi redBus sebagai representasi Akses Bus Kota di seluruh Indonesia seperti yang ditunjukkan pada Gambar 8.



Gambar 9. Wordcloud Sentimen Positif

Dapat disimpulkan bahwa 10 kata yang paling dominan pada sentimen positif ulasan pengguna aplikasi redBus sebagai representasi Akses Bus Kota di seluruh Indonesia meliputi kata terima kasih; baik; mantap; baik; cepat; sopir; oke; murah; baik; dan ramah. Visualisasi kata pada data ulasan sentimen netral ditunjukkan pada Gambar 9.



Gambar 10. Wordcloud Sentimen Netral

Kata-kata yang mendominasi sentimen netral pengguna aplikasi redBus sebagai representasi pengguna Access by Bus Kota di seluruh Indonesia antara lain adalah kata-kata aplikasi; pengemudi; puas; peta; daftar; bagus; pesanan; pengemudi; dan harga. Visualisasi kata-kata pada data ulasan sentimen negatif ditunjukkan pada Gambar 10.



Gambar 11. Wordcloud Sentimen Negatif

Mengacu pada Gambar 5 dapat disimpulkan bahwa kata-kata yang mendominasi sentimen negatif adalah kata-kata driver; aplikasi; bus; order; lama; tolong; error; dan gagal.

Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan dapat disimpulkan bahwa kinerja terbaik pada percobaan dengan pembagian data latih dan data uji, serta berbagai nilai k diperoleh pada percobaan dengan pembagian 90% data latih, 10% data uji dan menggunakan nilai $k = 5$ dengan nilai akurasi, presisi, dan recall secara berurutan sebesar 90,23%; 90,23%; dan nilai recall sebesar 72,38%. Klasifikasi sentimen dengan model terbaik menggunakan parameter $k = 3$ menghasilkan 79,26% sentimen positif, 17,25% sentimen netral, dan 3,49% sentimen negatif.

Ucapan Terima Kasih

Berisi deskripsi tentang ucapan terima kasih atau pengakuan kepada pihak-pihak (perseorangan atau institusi) yang turut terlibat secara langsung atau tidak langsung dalam mensukseskan kegiatan program penelitian yang telah dilakukan. (Times New Roman 10 point dengan spasi tunggal)

Daftar Pustaka

- [1] Abd, D. H., Abbas, A. R., & Sadiq, A. T. (2021). Analyzing sentiment system to specify polarity by lexicon-based. *Bulletin of Electrical Engineering and Informatics*, 10(1), 283–289. <https://doi.org/10.11591/eei.v10i1.2471>
- [2] Alnuaimi, A. F. A. H., & Albaldawi, T. H. K. (2024). An Overview of Machine Learning Classification Techniques. *BIO Web of Conferences*, 97(00133), 1–24. <https://doi.org/10.1051/bioconf/20249700133>
- [3] Hakim, Z. R., & Sugiyono, S. (2024). Analisa Sentimen Terhadap Kereta Cepat Jakarta – Bandung Menggunakan Algoritma Naïve Bayes Dan K-Nearest Neighbor. *Jurnal Sains Dan Teknologi*, 5(3), 939–945. <https://doi.org/10.55338/saintek.v5i3.1423>
- [4] Harahap, E. P., Purnomo, H. D., Iriani, A., Sembiring, I., & Nurtino, T. (2024). Trends in Sentiment of Twitter Users Towards Indonesian Tourism: Analysis with The K-Nearest Neighbor Method. *Computer Science and Information Technologies*, 5(1), 19–28. <https://doi.org/10.11591/csit.v5i1.pp19-28>
- [5] Homaidi, A., & Fatah, Z. (2024). Implementasi Metode K-Nearest Neighbors (KNN) untuk Klasifikasi Penyakit Jantung. *G-Tech: Jurnal Teknologi Terapan*, 8(3), 1720–1728.
- [6] Irawan, F. R., Jazuli, A., & Khotimah, T. (2022). Analisis Sentimen Terhadap Pengguna Gojek Menggunakan Metode K-Nearset Neighbors. *JIKO (Jurnal Informatika Dan Komputer)*, 5(1), 62–68. <https://doi.org/10.33387/jiko>
- [7] Iwandini, I., Triayudi, A., & Soepriyono, G. (2023). Analisa Sentimen Pengguna Transportasi Jakarta Terhadap Transjakarta Menggunakan Metode Naives Bayes dan K-Nearest Neighbor. *Journal of Information System Research (JOSH)*, 4(2), 543–550. <https://doi.org/10.47065/josh.v4i2.2937>
- [7] Jaman, J. H., Abdulrohman, R., Suharso, A., Sulistiowati, N., & Dewi, I. P. (2020). Sentiment Analysis on Utilizing Online Transportation of Indonesian Customers Using Tweets in The Normal Era and The Pandemic Covid-19 Era with Support Vector Machine. *Advances in Science, Technology and Engineering Systems Journal*, 5(5), 389–394. <https://doi.org/10.25046/AJ050549>
- [8] Javid, M. A., Ali, N., Hussain Shah, S. A., & Abdullah, M. (2021). Travelers' Attitudes Toward Mobile Application-Based Public Transport Services in Lahore. *SAGE Open*, 11(1), 1–15. <https://doi.org/10.1177/2158244020988709>
- [9] Liza, B., Utami, A., Asyidiqi, H., & Rizki, K. Z. (2024). Analysis of Gojek ' s Expansion Strategy to Vietnam in 2018-2022. *Globalization & International Relations*, 1(1), 44–71.
- [10] Mogaji, E., & Nguyen, N. P. (2024). Evaluating the emergence of contactless digital payment technology for transportation. *Technological Forecasting and Social Change*, 203(123378), 1–18. <https://doi.org/10.1016/j.techfore.2024.123378>
- [11] Mois, G., & Rogers, W. A. (2024). Developing a framework for digital activities of daily living. *The Gerontologist*, 64(110), 1–8. <https://doi.org/10.1093/geront/gnae110>
- [12] Mouratidis, K., Peters, S., & van Wee, B. (2021). Transportation technologies, sharing economy, and teleactivities: Implications for built environment and travel. *Transportation Research Part D: Transport and Environment*, 92(102716), 1–23. <https://doi.org/10.1016/j.trd.2021.102716>
- [13] Natasya, D., & Fatah, Z. (2024). *Gudang Jurnal Multidisiplin Ilmu Penerapan Data Mining Untuk Analisis Data Sosial Media Menggunakan Metode K-Nearest Neoghbor (K-NN)*. 2(November), 105–109.
- [13] Nuraini, Putra, P. P., & Chan, A. S. (2024). Cosmetic Product Review Sentiment Using K-Nearest Neighbor. *International Journal of Research in Vocational Studies (IJRVOCAS)*, 3(4), 150–158. <https://doi.org/10.53893/ijrvocas.v3i4.83>
- [14] Oladimeji, D., Gupta, K., Kose, N. A., Gundogan, K., Ge, L., & Liang, F. (2023). Smart Transportation: An Overview of Technologies and Applications. *Sensors*, 23(8), 1–32.

<https://doi.org/10.3390/s23083880>

- [15] Patil, R. N., Singh, Y. P., Rawandale, S. A., & Singh, S. (2024). Improving Sentiment Classification on Restaurant Reviews Using Deep Learning Models. *Procedia Computer Science*, 235, 3246–3256. <https://doi.org/10.1016/j.procs.2024.04.307>
- [16] Prakoso, R. (2017). KBBA. *Github.Com*.
- [17] Prananda, N. I., Chan, B. Y., Lim, L. C., & Poon, W. C. (2020). Go-Jek Company : Go-Jek's Rise to Dominating Indonesian's Markets and Southeast Asean. *TEST Engineering and Management*, 82(1), 735–743.
- [18] Sutrisno, E. P., & Amini, S. (2023). *IMPLEMENTASI ALGORITMA K-NEAREST NEIGHBOR PADA IMPLEMENTATION OF K-NEAREST NEIGHBOR ALGORITHM IN SENTIMENT ANALYSIS OF USER REVIEWS FOR DIGITAL*. 2(April 2021), 687–695.
- [19] Volti, R., & Croissant, J. L. (2024). *Society & Technological Change* (Ninth Edit). Waveland Press Inc.
- [20] Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. In *Artificial Intelligence Review* (Vol. 55, Issue 7). Springer Netherlands. <https://doi.org/10.1007/s10462-022-10144-1>
- [21] Yang, J. (2024). Rethinking Tokenization: Crafting Better Tokenizers for Large Language Models. *International Journal of Chinese Linguistics*, 11(1), 94–109.