Bit (Fakultas Teknologi Informasi Universitas Budi Luhur) p-ISSN: 1693-9166
Vol. 21, No. 2, September 2024, hlm. 161-168 e-ISSN: 2685-127X

ANALISIS SENTIMEN ULASAN PENGGUNA GOJEK DAN INDRIVE PADA GOOGLE PLAYSTORE DENGAN ALGORITMA K-NEAREST NEIGHBOR

Teguh Febriyanto [1*], Achmad Solichin[2], Windhy Widhyanty[3]

[1,2]Teknik Informatika, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta Selatan, Indonesia

Email: [1*]tfebriyan10@gmail.com, [2]achmad.solichin@budiluhur.ac.id, [3]windhy.widhyanty@budiluhur.ac.id

(Naskah masuk: 7 Agustus 2024, diterima untuk diterbitkan: 7 September 2024)

Abstrak

Transportasi adalah sarana atau suatu hal kendaraaan yang umum digunakan sebagian masyarakat di Indonesia untuk melakukan kegiatan memindahkan barang ataupun manusia dari suatu tempat ke tempat yang lainnya. Transportasi sangat diperlukan dalam kegiatan sehari-hari namun seiring berjalannya waktu, banyak penumpukan alat transportasi di kota-kota besar seperti kota metropolitan Jakarta, tidak jarang juga ditemukan penyebab kecelakaan disebabkan terlalu banyak transportasi di suatu daerah. Maka pemerintah melakukan penghimbauan agar masyarakat terutama di daerah perkotaan padat penduduk menggunakan transportasi umum. Transportasi umum itu banyak jenisnya dari angkutan umum hingga transportasi online yang lagi trend masa kini, banyak developer maupun perusahaan dari dalam negeri maupun luar negeri. Dengan banyaknya pilihan transportasi online, dibutuhkan data ulasan pengguna yang digunakan untuk melihat tingkat kepuasaan atau komentar dari pengguna dari aplikasi tersebut semisal dari gojek dan indrive. Oleh karena itu dilakukan analisis sentimen dari data ulasan pengguna aplikasi gojek dan indrive yang diambil melalui platform google playstore yang diambil pada bulan juni 2024 menggunakan google playstore scraping. Dari data kotor yang didapat, data tersebut di bersihkan melalui proses preprocessing lalu diberi label dengan menggunakan lexicon based setelah itu dilakukan split data yang menghasilkan data training dan juga data testing. Setelah di dapatkan data training lalu di hitung bobot perkata dengan TF-IDF dan nantinya data testing akan diklasifikasikan dengan metode algoritma K-Nearest Neighbor (KNN) menjadi dua label terpisah yaitu positif dan negatif. Bahwa dalam pengujian menggunakan K-Nearest Neighbor (KNN), dengan menggunakan k yang sama yaitu k=3, didapatkan akurasi pada dataset aplikasi gojek yaitu 72% dengan menggunakan rasio data 80:20 dari 860 data, sedangkan didapatkan akurasi pada dataset aplikasi indrive yaitu 85% dengan menggunakan rasio 80:20 dari 1477 data.

Kata kunci: *Gojek, Indrive, Analisis Sentimen, K-Nearest Neighbor (KNN), Text mining*

SENTIMENT ANALYSIS OF USER REVIEWS FOR GOJEK AND INDRIVE ON GOOGLE PLAYSTORE USING THE K-NEAREST NEIGHBOR ALGORITHM

Abstract

Transportation is a means or a vehicle that is commonly used by some people in Indonesia to carry out activities to move goods or people from one place to another. Transportation is very necessary in daily activities but over time, there is a lot of accumulation of transportation equipment in big cities such as the metropolitan city of Jakarta, it is not uncommon to find the cause of accidents due to too much transportation in an area. So the government makes an appeal so that people, especially in densely populated urban areas, use public transportation. There are many types of public transportation from public transportation to online transportation which is the current trend, many developers and companies from within the country and abroad. With so many online transportation options, user review data is needed which is used to see the level of satisfaction or comments from users of these applications such as gojek and indrive. Therefore, sentiment analysis is carried out from the gojek and indrive application user review data taken through the google playstore platform which was taken in June 2024 using google playstore scraping. From the dirty data obtained, the data is cleaned through a preprocessing process and then labeled using lexicon based after that split data is carried out which results in training data and also testing data. After getting the training data, the weight of each word is calculated with TF-IDF and laer the testing data will be classified by the K-Nearest Neighbor (KNN) algorithm method into two separate labels, namely positive and negative. That in testing using K-Nearest Neighbor (KNN), using the same k,

namely k = 3, the accuracy obtained on the gojek application dataset is 72% using a data ratio of 80: 20 of 860 data, while the accuracy obtained on the indrive application dataset is 85% using a ratio of 80: 20 of 1477 data.

***Keywords**: Gojek, Indrive, Sentiment Analysis, K-Nearest Neighbor (KNN), Text mining*

1. PENDAHULUAN

Transportasi adalah sarana atau suatu hal kendaraaan yang umum digunakan sebagian masyarakat di Indonesia untuk melakukan kegiatan memindahkan barang ataupun manusia dari suatu tempat ke tempat yang lainnya [1]. Transportasi sangat diperlukan dalam kegiatan sehari-hari namun seiring berjalannya waktu, banyak penumpukan alat transportasi di kota-kota besar seperti kota metropolitan Jakarta, tidak jarang juga ditemukan penyebab kecelakaan disebabkan terlalu banyak transportasi di suatu daerah. Maka pemerintah melakukan penghimbauan agar masyarakat terutama di daerah perkotaan padat penduduk menggunakan transportasi umum.

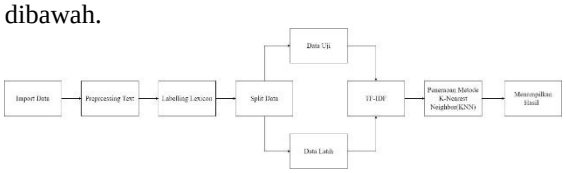
Oleh sebab itu banyak dinamika yang membuat kebingungan memilih berbagai macam aplikasi ojek online yang dapat digunakan dalam kegiatan sehari-hari. Maka dari itu dibutuhkan suatu sistem yang dapat mengklasifikasi sentimen ulasan seseorang otomatis dalam kelas positif dan negatif yang nantinya akan menampilkan beberapa saran dari ulasan pengguna lainnya yang akan ditujukan untuk pengguna yang bingung memilih aplikasi ojek online apa saja. Google play store sebagai wadah yang menjadi tempat pengunduhan aplikasi ini memiliki fitur yang cukup bermanfaat untuk melihat ulasan-ulasan pengguna aplikasi tersebut. Fungsi awal dari adanya ulasan pengguna ini bisa dimanfaatkan sebagai tolak ukur yang efektif dan juga efisien untuk menemukan informasi yang valid pada suatu aplikasi tertentu. Ulasan yang muncul berupa saran positif dan negatif dari suatu aplikasi tersebut. Untuk melihat dan menyortir beberapa ulasan tersebut secara manual cukup sulit dilakukan. Maka ada sebuah metode yang digunakan untuk memudahkan mekalukan pekerjaan tersebut.

Analisis Sentimen adalah teknik yang digunakan. Ini adalah proses pengolahan data teks yang menilai apakah pendapat dalam kalimat positif atau negatif [2]. Banyak algoritma yang terkait dengan analisis sentiman, salah satunya adalah Algoritma K-Nearest Neighbour (KNN). Algoritma ini adalah metode klasifikasi yang mengelompokkan data ke dalam kelas-kelas yang telah ditentukan sebelumnya berdasarkan jarak terdekat, atau nilai K yang biasanya disebut [3]. Studi sebelumnya tentang analisis sentimen termasuk penelitian yang mengevaluasi penggunaan algoritma K-Nearest Neighbor untuk mengkategorikan ulasan pengguna dari aplikasi Digital Korlantas POLRI. Penelitian ini menggunakan 600 data ulasan pengguna, yang dibagi menjadi 70% untuk data latihan dan 30% untuk data uji. Nilai K dicari dengan cosine similarity, dengan nilai akurasi tertinggi k=9 [4]. Studi sebelumnya yang menggunakan algoritma naive bayes untuk menganalisis sentimen ulasan pengguna untuk aplikasi Mypertamina di Google Playstore, yang menggunakan 3498 data, menemukan bahwa penelitian tersebut memiliki nilai akurasi sebesar 91%, presisi sebesar 92%, dan recall sebesar 100% [5].

Fokus dari penelitian ini adalah melakukan sentimen analisis tentang dua perusahaan ojek online yang berada di Indonesia berdasarkan ulasan pengguna yang tersedia di Google Playstore menggunakan metode K-Nearest Neighbor (KNN). Penggunaan algoritma K-nearest Neighbor (KNN) dengan menggunakan fitur TF-IDF ini diharapkan dapat mengklasifikasikan ulasan pengguna yang berada di kolom komentar Google Playstore dengan baik, sehingga akan menghasilkan klasifikasi yaitu ulasan positif dan negatif. Hasil pengujian akan dilakukan menggunakan confusion matrix, dan nantinya sistem analisis sentimen ini akan berbentuk aplikasi berbasis web dan metode yang digunakan untuk mengumpulkan data dalam penelitian ini adalah web scraping.

2. METODE PENELITIAN

Dalam penelitian ini yang dikembangkan adalah sistem informasi yang menggunakan implementasi metode K-Nearest Neighbor(KNN). Dimana dalam sistem tersebut ada tahapan-tahapan yaitu import data, labeling data, preprocessing data, dan juga perhitungan TFIDF beserta perhitungan prediksi dari metode K-Nearest Neighbor. Tahapan metode penelitian tersebut dapat di lihat pada Gambar 1 dibawah.



Gambar 1. Metode Penelitian

2.1 Pengumpulan Data

Data ulasan yang didapatkan dari platform Google Play Store ini diambil dari komentar ulasan pengguna dari aplikasi Gojek dan Indrive dengan teknik scraping menggunakan library google colab. Yang dilakukan pada tahapan ini adalah menarik ulasan Gojek dan Indrive di platform Google Playstore ke dalam file berbentuk CSV lalu disimpan ke dalam database sebagai data kotor yang nantinya akan diproses dan digunakan untuk penelitian kali ini dan dapat di implementasikan ke dalam algoritma K-NN untuk analisis sentimen. Data yang didapatkan dalam ulasan Gojek sekitar 1500 data kotor.

2.2 Preprocessing Data

Text preprocessing atau prapemrosesan teks adalah langkah awal dalam melakukan text mining. Prapemrosesan teks biasanya melibatkan penghapusan data yang tidak relevan atau transformasi data teks menjadi format yang lebih mudah di proses oleh sistem [6]. Dalam preprocessing ini terdapat beberapa tahapan yaitu case folding, cleansing, slangword, stopword removal, tokenizing, stemming.

a. *Case folding* yaitu suatu proses yang dilakukan untuk mengubah semua huruf yang berbentuk kapital dalam suatu kalimat menjadi huruf yang berbentuk kecil (lowercase).
b. *Cleansing* suatu proses yang dilakukan untuk menghilangkan beberapa karakter yag tidak dibutuhkan di dalam dokumen.
c. *Slangword* yaitu suatu proses yang dilakukan untuk mengubah atau menormalisasikan kata yang dilakukan dengan memperbaiki kata-kata yang disingkat menjadi kata yang memiliki arti sama, sesuai yang ada di dalam KBBI.
d. *Stopword* removal yaitu suatu proses yang dilakukan untuk mengambil kata-kata yang penting dan membuang kata-kata yang kurang penting dalam suatu dokumen. Pada tahapan ini teks sebelum dilabeli akan di hilangkan terlebih dahulu teks yang tidak ada hubungannya dengan analisis sentimen sehingga besarnya teks akan berkurang tanpa mengurangi isi sentimen teks [7].
e. *Tokenizing* yaitu suatu proses pemisahan atau pemecahan dokumen yang asalnya dari suatu kalimat menjadi potongan kata berdasarkan spasi, sehingga menghasilkan term-term yang terpisah.
f. *Stemming* yaitu suatu proses yang dilakukan untuk mengolah kata-kata yang memiliki imbuhan menjadi kata dasar dengan aturan-aturan tertentu [7].

2.3 Labelisasasi Kamus *Lexicon*

Labeling adalah suatu proses yang memberikan klasifikasi berdasarkan karakteristik yang ada di dalam sebuah kalimat dalam suatu dokumen[8]. Pelabelan yang dilakukan atau biasa disebut pemberian polaritas pada dataset yang sudah bersihkan dalam tahap preprocessing. Pengelompokkan polaritas dilakukan dengan cara otomatis menggunakan lesxicon dengan bobot polaritas yang berbeda. Nilai positif dinyatakan 1-5 bobot, sedangkan -1 sampai dengan -5 dinyatakan sebagai nilai negatif. [9].

Dalam proses pelabelan, leksikon dengan kamus tambahan digunakan untuk menentukan kata-kata yang digunakan. Untuk menghitung polaritas, bobot kata dalam setiap komentar dijumlahkan. Hasilnya menunjukkan apakah polaritas positif atau negatif. Komentar dianggap positif jika polaritas positif, dan komentar dianggap negatif jika polaritas negatif.

2.4 TF-IDF

Term Frequency-Inverce Document Frequency(TF-IDF) adalah suatu metode pembobotan yang dilakukan untuk menghitung pada setiap bobot kata yang terdapat dalam suatu dokumen yang tersedia. Dokumen-dokumen tersebut yang nantinya akan dirubah menjadi vektor dengan jumlah kata (term) yang digunakan untuk pengklasifikasian[10]. Berikut rumus TF-IDF:

a. *Term Frequency* (TF)

$$tf_{(kd)} = \frac{jumlah\ frekuensi\ istilah\ k\ yang\ muncul\ dalam\ dokumen\ d}{jumlah\ istilah\ dalam\ dokumen} \quad (1)$$

b. *Inverce Document Frequency* (IDF)

$$idf_{(k)} = log\frac{N}{df_k} \quad (2)$$

c. TF-IDF

$$tfidf_{(k,d)} = tf_{(k,d)} * idf_{(k)} \quad (3)$$

Keterangan :

$tf_{(kd)}$ = Frekuensi kemunculan kata (*term*) k dalam dokumen d
k = Suatu kata (*term*)
d = Suatu dokumen
$idf_{(k)}$ = *Inverse Document Frequency* k
N = Jumlah dokumen yang ada di dalam *database*
df_k = Jumlah dokumen yang mengandung kata (*term*) k
$tf - idf_{(k,d)}$ = Bobot kata (*term*) k terhadap dokumen d

2.5 Algoritma K-NN

Algoritma K-Nearest Neighbor (KNN) adalah suatu metode yang termasuk di dalam supervised learning dan melibatkan proses pembelajaran dari data training [4]. menjelaskan cara kerja dasar dari metode KNN adalah mencari jarak terpendek antara data yang akan direklasifikasi dengan data latih yang ada dalam lingkungan k (jarak minimum). Data yang memiliki jarak terkecil dengan kategori/kelas terbesar akan menjadi kategori/kelas baru untuk data uji (testing data). Pada penelitian ini menggunakan perhitungan jarak Euclidean Distance [11] yang ditunjukkan pada persamaan.

2.6 Confusion Matrix

Confusion matrix adalah sebuah metode yang seringkali dimanfaatkan dalam penelitian untuk mengetahui atau evaluasi dari hasil dan pengukuran performa dalam suatu metode klasifikasi. Dalam hal ini penggunaan confusion matrix sangatlak penting dikarenakan untuk mengukur sejauh mana sistem yang dibuat dapat melakukan proses klasifikasi data dengan baik. Pada confusion matrix, ada beberapa perhitungan yang akan dilakukan yaitu akurasi, presisi, dan recall. Dibawah ini adalah perhitungan akurasi, precision dan recall [12] yang dapat dilihat pada persamaan sebagai berikut:

$$accuracy = \frac{TP+TN}{TP+FN+FP+TN} X\ 100\% \quad ... (5)$$

$$precision = \frac{TP}{TP+FP} X\ 100\% \quad \ldots. \quad (6)$$

$$recall = \frac{TP}{TP+FN} X\ 100\% \quad \ldots \quad (7)$$

$$f1score = 2\ X \frac{Precision\ X\ Recall}{Precision+Recall} X\ 100\% \quad \ldots. (8)$$

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Data yang dimasukkan dalam peneitian kali ini di dapat dari ulasan aplikasi Gojek dan juga Indrive pada Google Play Store dan selanjutnya data di import kedalam database sistem. Tabel 1 menyajikan contoh dari data ulasan yang diambil pada aplikasi Gojek dan Indrive di Google Play Store.

Tabel 1. Tabel Komentar Aplikasi Indrive

No	*Username*	*Content*	*Score*
1	Naila Square	auto bid nya di aktifkan lagi dong	5
2	Dian Abdul Hakim	Parah malah ada peringatan padahal pake aplikasi ori dari awal juga	1
3	Hana Priyandi	gila yaaaa fungsi autobit itu buat Nerima orderan secara otomatis kn kenapa ga berfungsi sih kalo buat sistem yg bener dong makin kalah aja sama apk sebelah hadeeeeh indriver indriver kacau lu	1
…	…	…	…
1,993	Terry Irwansyah	Mayan,buat nambah penghasilan	5

Tabel 2. Tabel Komentar Aplikasi Gojek

No	*Username*	*Content*	*Score*
1	Rachel mutiara adita	Ongkirnya g ngotak mahalnya padahal dekat loh jaraknya	1
2	Ardi Budiman	kadang klo lagu hujan driver nya jauh2	2
3	Nizar Ramadhan	saldo ga balik padahal di sana tertera gagal	1
…	…	…	…
1,993	Tia Lovely	Go-Jek banyak diskonnya mantaap	5

3.2 Preprocessing Data

Tahap selanjutnya yaitu melakukan pembersihan data dimana data tersebut masih dalam keadaan kotor, lalu dibersihkan menjadi cleantext yang nantinya dapat di proses untuk perhitungan TF-IDF dan juga klasifikasi KNN. berikut adalah tahapan preprocessing yang dilakukan dalam penelitian kali ini:

a. *Case Folding*

Proses *case folding* yaitu melakukan perubahan huruf kapital menjadi huruf kecil pada selurut data yang ada di dalam dokumen tersebut. Proses *case folding* dapat di lihat pada tabel 3.

Tabel 3. Case Folding

Content	*Casefolding*
Iklan Gojek SANGAT MENGGANGGU SAAT MAIN GAME...!!!!!!! TIAP MENIT IKLAN GOJEK TERUS!!!	iklan gojek sangat mengganggu saat main game...!!!!!!! tiap menit iklan gojek terus!!!

b. *Cleansing*

Proses cleansing yaitu suatu tahapan yang berguna untuk menghilangkan karakter atau simbol-simbol yang tidak diperlukan pada data ulasan pengguna, termasuk tanda baca, angka, URL, dan juga simbol („”~&?!><#%{}([0-9]+:;‟). Proses cleansing dapat di lihat pada tabel 4.

Tabel 4. Cleansing

Casefolding	Cleansing
Kenapa promo cuma paylater? Dipaksa nyicil? Ngutang?	kenapa promo cuma paylater dipaksa nyicil ngutang

c. Tokenisasi

Proses tokenisasi yaitu suatu tahapan yang dilakukan untuk memecahkan sebuah kalimat menjadi potongan-potongan kata yang berdiri sendiri, proses ini dilakukan agar mempermudah untuk melakukan proses *slangword* dan juga *stopword*. Proses tokenisasi dapat di lihat pada tabel 5.

Tabel 5. Tokenisasi

Cleansing	Tokenisasi
kenapa promo cuma *paylater*? dipaksa nyicil? ngutang?	["kenapa", "promo", "cuma", "*paylater*", "dipaksa", "nyicil", "ngutang"]

d. *Slangword*

Proses *slangword* yaitu suatu tahapan yang berguna untuk mengubah atau mengganti kata-kata yang disingkat menjadi kata-kata yang terdaftar di Kamus Besar Bahasa Indonesia (KBBI) dengan arti sama tanpa merubah maksud darti kata tersebut. Proses *slangword* dapat di lihat pada tabel 6.

Tabel 6 . Slangword

Tokenisasi	*Slangword*
aplikasinya tidak dapat dibuka, padahal sdh sy *update*, mhn ijin dibantu admin, trims	aplikasinya tidak dapat dibuka, padahal sudah saya *update*, mohon ijin dibantu admin, terimakasih

e. *Stopword Removal*

Stopword removal yaitu sebuah proses yang dilakukan untuk menghilangkan kata-kata yang tidak diperlukan atau tidak penting selama proses preprocessing. Proses *slangword* dapat di lihat pada tabel 7.

Tabel 7 Stopword Removal

Slangword	*Stopword Removal*
pelayanan ramah dan sopan	pelayanan ramah sopan

f. *Stemming*

Stemming adalah sebuah proses yang dilakukan untuk mengubah kata-kata yang memiliki awalan atau akhiran yang termasuk kata sambung atau kata imbuhan menjadi kata dasar di dalam Kamus Besar Bahasa Indonesia (KBBI). Proses *slangword* dapat di lihat pada tabel 8.

Tabel 8. Stemming

Stopword Removal	*Stemming*
sangat banget membantu	sangat banget bantu

3.3 Labelisasi Kamus *Lexicon*

Pembagian label pada komentar dilakukan untuk memberikan emosi kepada komentar sehingga pesan dan kesan dapat diidentifikasi. Positif adalah sentuhan yang menunjukkan dukungan atau kesetujuan, sedangkan negatif adalah sentuhan yang menunjukkan ketidakpuasan dari komentar padangan. Sentimen yang tidak menunjukkan aspek positif atau negatif disebut neutral. Untuk mengetahui perasaan dari komentar, proses pelabelan ini menggunakan leksikon.

Tabel 9. Labelisasi kamus *lexicon*

Document_id	*Cleantext*	*Label*	Skor
1	terima kasih selalu beri nyaman baik	*Positive*	7
2	moga promo diskon makin banyak biaya ongkos kirim biaya lain lebih murah	*Positive*	3
3	saat sedang cari driver belum sangkut ke salah satu *driver* tidak ada pilih batal padahal sudah tunggu lama	*Negative*	-21

Pada tahapan labelisasi kali ini menggunakan kamus *lexicon*, yang didapat dari github bernama Fajri91, lalu digunakan untuk melabelisasi kalimat di dokumen tersebut adalah positif dan negatif, sehingga pada saat splitdata nantinya akan menggunakan hasil data teks yang sudah dilabelisasi oleh kamus *lexicon*. Proses *labeling* dapat dilihat pada Tabel 9.

3.4 *Split* Data

Pada tahapan pembagian data, data yang sudah di beri label dengan kamus lexicon akan dilakukan proses pembagian data menjadi data training dan juga data testing. Pada penelitian ini dilakukan pembagian data dengan rasio 80:20.

Tabel 10. *Split* Data 80:20

Jenis Data	Jumlah
Data Training	688
Data Testing	172
Jumlah	860

3.5 Perhitungan Bobot TF-IDF

TF-IDF adalah suatu tahapan perhitungan bobot mengunakan metode Term Frequency-Inverse Document Frequency (TF-IDF). Pada tahapan ini melibatkan perhitungan nilai TF untuk setiap kata yang muncul setiap di dokumen tersebut. Setelah itu nilai TF di kalikan dengan nilai IDF, yang menggambarkan pentingnya kata tersebut dalam seluruh dokumen yang tersedia. Berikut adalah simulai perhitungan bobot TF-IDF.

Tabel 11. Data *Training*

Data Ulasan	Label
sarana transportasi bagus mahal	Positif
Data Ulasan	**Label**
aplikasi jelek masuk akun limit habis suruh tunggu jam	Negatif
jelek ah gocar hemat cancel melulu	Negatif
Gojek the best	Positif
pesan gocar jarak tidak mau ambil driver tunggu jam tidak ada ambil parah sih	Negatif

a. Menghitung *Term Frequency* (TF)

Setelah data yang sudah di preprocessing menjadi data bersih lalu langkah selanjutnya adalah perhitungan TF untuk setiap kata di dalam suatu dokumen. Perhitungan *Term Frequency* (TF) yang terdapat di Tabel 12.

b. Menghitung *Inverse Document Frequency* (IDF)

Pada perhitungan *Inverse Document Frequency* (IDF) proses ini menggunakan suatu term yang terdapat dalam seluruh dokumen. Untuk mendapatkan nilai IDF adalah dengan menghitung jumlah total dokumen (N) akan dibagi dengan jumlah DF, lalu nilai IDF di dapat dari log hasil perhitungan tersebut. Simulasi perhitungan IDF terdapat di tabel 13.

c. TF-IDF

Setelah dilakukan perhitungan IDF, selanjutnya adalah melakukan perhitungan TF-IDF yang didapatkan dari mengalikan nilai TF dengan nilai IDF yang terlah dijelaskan sebelumnya. Simulasi perhitungan TF-IDF dapat dilihat pada tabel 14.

Tabel 12. Menghitung *Document Frequency*

Kata *Term*	D1	D2	D3	D4	D5
sarana	1				
transportasi	1				
bagus	1				
mahal	1				
aplikasi		1			
jelek		1	1		
masuk		1			
akun		1			
limit		1			
habis		1			
suruh		1			
tunggu		1			1
jam		1			1
ah			1		
gocar			1		1
hemat			1		
cancel			1		
melulu			1		
gojek				1	
the				1	
best				1	
pesan					1
jarak					1
tidak					2
mau					1
ambil					2
driver					1
ada					1
parah					1
sih					1

Tabel 13. *Inverse Document Frequency*

Kata *Term*	DF	IDF
sarana	1	0,698970004
transportasi	1	0,698970004
bagus	1	0,698970004
mahal	1	0,698970004
aplikasi	1	0,698970004
jelek	2	0,397940009
masuk	1	0,698970004
akun	1	0,698970004
limit	1	0,698970004
habis	1	0,698970004
suruh	1	0,698970004
tunggu	2	0,397940009
jam	2	0,397940009
ah	1	0,698970004
gocar	2	0,397940009
hemat	1	0,698970004
cancel	1	0,698970004
melulu	1	0,698970004
gojek	1	0,698970004
the	1	0,698970004
best	1	0,698970004
pesan	1	0,698970004
jarak	1	0,698970004
tidak	1	0,698970004
mau	1	0,698970004
ambil	1	0,698970004
driver	1	0,698970004
ada	1	0,698970004
parah	1	0,698970004
sih	1	0,698970004

Tabel 14. TF-IDF

Kata *Term*	D1	D2	D3	D4	D5
sarana	0,69				
transportasi	0,69				
bagus	0,69				
mahal	0,69				
aplikasi		0,69			
jelek		0,69	0,69		
masuk		0,69			
akun		0,69			
limit		0,69			
habis		0,69			
suruh		0,69			
tunggu		0,69			0,39
jam		0,69			0,39
ah			0,69		
gocar			0,69		0,39
hemat			0,69		
cancel			0,69		
melulu			0,69		
gojek				0,69	
the				0,69	
best				0,69	
pesan					0,69
jarak					0,69
tidak					1,38
mau					0,69
ambil					1,38
driver					0,69
ada					0,69
parah					0,69
sih					0,69

3.6 Klasifikasi *K-Nearest Neighbor*

Setalah dilakukan pembobotan kata pada data ulasan aplikasi gojek dan indrive, Selanjutnya, klasifikasi dilakukan dengan metode K-Nearest Neighbor (KNN). Metode ini mengambil k-tetangga terdekat dengan menggunakan jarak Euclidean. Langkah-langkah dalam proses K-Nearest Neighbor (KNN) dibawah ini.

a. Sample Data *Testing*

Tabel 15. Data *Testing*

Data Ulasan	Label
tidak jelas ke sini eror *driver maps* gerak posisi driver aplikasi gojek jelek	Negatif

b. Menghitung Jarak *Euclidian*

$$d_{latih1,uji1} = \sqrt{\begin{matrix}(0{,}4761)+(0{,}4761)+(0{,}4761)+(0{,}4761)+(0{,}4761)+(0{,}1521)\\ +(0)+(0)+(0)+(0)+(0)+(0)+(0)+(0)+(0)+(0)+(0)\\ +(0)+(0{,}4761)+(0)+(0)+(0)+(0)+(0{,}4761)+(0)+(0)\\ +(1{,}9044)+(0)+(0)+(0)\end{matrix}}$$

$$d_{latih1,uji1} = \sqrt{5{,}3892}$$

$$d_{latih1,uji1} = 2{,}3214$$

Dari contoh sebelumnya, hasil perhitungan setiap jarak antara vektor data training dan juga *data testing*.

Tabel 16. Jarak *Euclidian*

Jarak *Euclidian*
d_(latih1,uji1) = 2,3214
d_(latih2,uji1) = 2,3856
d_(latih3,uji1) = 2,3521
d_(latih4,uji1) = 2,0166
d_(latih5,uji1) = 2,7337

c. Mencari K Tetangga Terdekat

Proses selanjutnya adalah mencari nilai K terdekat setelah melakukan perhitungan jarak antara data *training* dan juga data *testing*, menentukan jarak terdekat dengan cara diurutkan berdasarkan nilai yang terdekat dengan data *testing* yang sudah di hitung sebelumnya. Tabel berikut menunjukkan hasil klasifikasi yang diperoleh berdasarkan nilai k yang digunakan.

Tabel 17. Mencari K Tetangga

Urutan	Jarak *Euclidean*	Data ke -	*Label Data Training*
1	d_(latih4,uji1) = 2,0166	*Testing* 1, *Training* 4	Positif
2	d_(latih1,uji1) = 2,3214	*Testing* 1, *Training* 1	Positif
3	d_(latih3,uji1) = 2,3521	*Testing* 1, *Training* 3	Negatif

Di dapatkan dari k terdekat adalah *data training* 4, *training* 1, dan juga *training* 3. Jumlah perbandingan positif dan negatif yang didapatkan adalah terdapat 2 data dengan label positif dan 1 data dengan label negatif. Sehingga, hasil dari data testing pada pengujian ini memiliki tabel positif.

3.7 Confusion Matrix

Selanjutnya menampilkan Hasil yang di dapat dari klasifikasi KNN itu menghasilkan *Confusion Matrix*, pada gambar berikut adalah *confusion matrix*, akurasi, presisi, *recall*, dan *f1score* dari hasil klasifikasi KNN data gojek dan juga data *indrive*.

Tabel 18. *Confusion Matrix* Data Gojek

Akurasi	Presisi	*Recall*	F1*score*
72%	0%	0%	0%

Tabel 19. *Confusion Matrix* Data *Indrive*

Akurasi	Presisi	*Recall*	F1*score*
85%	0%	0%	0%

Perbandingan akurasi Indrive dan Gojek

Tabel 20. Perbandingan Hasil Akurasi, Presisi, *Recall*, dan f1-score

Aplikasi	Jumlah Data *Testing*	Akurasi	Presisi	*Recall*	F1*score*
Gojek	172	72%	0%	0%	0%
Indrive	295	85%	0%	0%	0%

Berdasarkan tabel 20 dapat ditarik kesimpulan bahwa dalam pengujian menggunakan K-*Nearest Neighbor* (KNN), dengan menggunakan k yang sama yaitu k=3, didapatkan akurasi pada dataset aplikasi gojek yaitu 72% dengan menggunakan rasio data 80:20 dari 860 data, sedangkan didapatkan akurasi pada dataset aplikasi indrive yaitu 85% dengan menggunakan rasio 80:20 dari 1477 data. Nilai presisi di kedua data tersebut didapatkan hasil 0% dikarenakan didalam perhitungan presisi harus terdapat nilai TP sedangkan TP di dalam sentimen Gojek dan *Indrive* tidak memiliki nilai klasifikasi TP. Begitu juga dengan Recall yang membutuhkan nilai TP untuk mendapatkan hasil *recall* tersebut. Jika F1*score* didapatkan dari pengalian presisi dan recall, jika didapatkan hasil presisi dan *recall* 0% maka f1*score* didapatkan 0% juga.

Didapatkan nilai akurasi dari ulasan Gojek yaitu 72% dari ulasan yang berupa komentar negatif dengan sisanya sekitar 28% berkomentar positif. Ulasan itu didapatkan dari *Google Playstore* lalu diberi labelisasi dengan kamus *lexicon*. Begitu juga dengan nilai ulasan akurasi Indrive yaitu 85% dari ulasan yang berupa komentar negatif dengan sisanya 15% berkomentar positif. Nilai akurasi diatas menjadi besar karena algoritma yang memprediksi kata tersebut sama dengan hasil labelisasi kamus *lexicon*. Hal ini menunjukkan performa sistem aplikasi semakin data yang digunakan banyak, maka semakin akurat hasil yang akan didapatkan, dan juga dalam dataset yang terdapat pada aplikasi gojek lebih banyak merujuk ke positif, dengan akurasi yang lebih kecil menandakan komentar positif masih banyak terdapat di dalam data ulasan pengguna gojek.

4. KESIMPULAN

Penelitian ini berfokus pada penerapan metode klasifikasi teks untuk menganalisis sentimen publik terhadap aplikasi Gojek dan InDrive pada platform *Google Play Store*. Data dikumpulkan dari *Google Play Store* pada bulan Juni 2024, diproses melalui praproses, pelabelan, pelatihan data, dan pengujian. Metode *K-Nearest Neighbor* digunakan untuk mengklasifikasikan data, mengevaluasi kinerja model klasifikasi. Matriks konfusi menunjukkan bahwa data yang digunakan lebih akurat, dengan 72% data negatif memiliki hasil lebih positif pada aplikasi Gojek dan 85% pada aplikasi *InDrive*. Penelitian ini juga menemukan bahwa jumlah data secara signifikan memengaruhi kinerja algoritma dan kinerja aplikasi. Saran yang dapat penulis berikan untuk pengembangan sistem ini lebih lanjut adalah sebagai berikut: Pengembang diharapkan untuk mempercantik kembali tampilan aplikasi agar lebih menarik. Selain itu, pengembang selanjutnya diharapkan dapat menambahkan lebih banyak metode untuk memungkinkan perbandingan tingkat performa aplikasi. Penambahan jumlah data yang sangat besar memang dapat mempengaruhi akurasi hasil, namun dengan bertambahnya data, pengembang diharapkan juga dapat menemukan solusi agar kinerja aplikasi tidak menjadi terlalu lambat.

DAFTAR PUSTAKA

[1] A. N. Hasanah dan B. N. Sari, "Analisis Sentimen Ulasan Pengguna Aplikasi Jasa Ojek Online Maxim Pada Google Play Dengan Metode Naïve Bayes Classifier," *J. Inform. dan Tek. Elektro Terap.*, vol. 12, no. 1, hal. 90–96, 2024, doi: 10.23960/jitet.v12i1.3628.

[2] A. Setiawan, "Perbandingan Penggunaan Jarak Manhattan, Jarak Euclid, dan Jarak Minkowski dalam Klasifikasi Menggunakan Metode KNN pada Data Iris," *J. Sains dan Edukasi Sains*, vol. 5, no. 1, hal. 28–37, 2022, doi: 10.24246/juses.v5i1p28-37.

[3] M. K. Anam, B. N. Pikir, dan M. B. Firdaus, "Penerapan Na¨ive Bayes Classifier, K-Nearest Neighbor (KNN) dan Decision Tree untuk Menganalisis Sentimen pada Interaksi Netizen danPemeritah," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 21, no. 1, hal. 139–150, 2021, doi: 10.30812/matrik.v21i1.1092.

[4] E. P. Sutrisno dan S. Amini, "Implementasi Algoritma K-Nearest Neighbor Pada Analisis Sentimen Ulasan Pengguna Aplikasi Digital Korlantas Polri," *Pros. Semin. Nas. Mhs. Fak. Teknol. Inf.*, vol. 2, no. 2, hal. 687–695, 2023.

[5] G. Darmawan, S. Alam, dan M. I. Sulistyo, "Analisis Sentimen Berdasarkan Ulasan Pengguna Aplikasi Mypertamina Pada Google Playstore Menggunakan Metode Naïve Bayes," *STORAGE – J. Ilm. Tek. dan Ilmu Komput.*, vol. 2, no. 3, hal. 100–108, 2023.

[6] utomo budiyanto marlina hidayat, "SENTIMEN ANALISIS TENTANG HILIRISASI INDUSTRI BERDASARKAN OPINI MASYARAKAT DI TWITTER MENGGUNAKAN METODE K-NEAREST NEIGHBOR," vol. 2, no. September, hal. 826–835, 2023, [Daring]. Tersedia pada: http://3.8.6.95/ijcs/index.php/ijcs/article/view/3154%0Ahttp://3.8.6.95/ijcs/index.php/ijcs/article/download/3154/104.

[7] M. F. Rizki, W. Pramusinto, M. Hardjianto, dan S. Subandi, "Implementasi Algoritma K-Nearest Neigbors Untuk Analisis Sentimen Aplikasi Jobstreet," *Pros. Semin. Nas. Mhs. Fak. Teknol. Inf.*, vol. 2, no. 1, hal. 267–276, 2023.

[8] M. Priandi dan Painem, "Analisis Sentimen Masyarakat Terhadap Pembelajaran Daring di Era Pandemi Covid-19 pada Media Sosial Twitter Menggunakan Ekstraksi Fitur Countvectorizer dan Algoritma K-Nearest Neighbor," *Semin. Nas. Mhs. Ilmu Komput. dan Apl.*, vol. 2, no. 2, hal. 311–319, 2021.

[9] S. Mujahidin, B. Prasetio, dan M. C. C. Utomo, "Implementasi Analisis Sentimen Masyarakat Mengenai Kenaikan Harga BBM Pada Komentar Youtube Dengan Metode Gaussian naïve bayes," *Voteteknika (Vocational Tek. Elektron. dan Inform.*, vol. 10, no. 3, hal. 17, 2022, doi: 10.24036/voteteknika.v10i3.118299.

[10] M. Furqan, S. Sriani, dan S. M. Sari, "Analisis Sentimen Menggunakan K-Nearest Neighbor Terhadap New Normal Masa Covid-19 Di Indonesia," *Techno.Com*, vol. 21, no. 1, hal. 51–60, 2022, doi: 10.33633/tc.v21i1.5446.

[11] A. Yudhana, S. Sunardi, dan A. J. S. Hartanta, "Algoritma K-Nn Dengan Euclidean Distance Untuk Prediksi Hasil Penggergajian Kayu Sengon," *Transmisi*, vol. 22, no. 4, hal. 123–129, 2020, doi: 10.14710/transmisi.22.4.123-129.

[12] S. Juniarsih, E. F. Ripanti, dan E. E. Pratama, "Implementasi Naive Bayes Classifier pada Opinion Mining Berdasarkan Tweets Masyarakat Terkait Kinerja Presiden dalam Aspek Ekonomi," *J. Sist. dan Teknol. Inf.*, vol. 8, no. 3, hal. 239, 2020, doi: 10.26418/justin.v8i3.39118.