Universitas Ahmad Dahlan

Jurnal Ilmiah Matematika
Vol. 8, No. 2, October 2021, pp. 84-97
ISSN 2774-3241
http://journal.uad.ac.id/index.php/Konvergensi

Perbandingan 5 Jarak K-Nearest Neighbor pada Analisis Sentimen

Almuzhidul Mujhid [a,1,*], Aris Thobirin [a,2], Salma Nadya Firdausy [a,3], Sugiyarto [a,4], Lanova Ade Rahmadani [a,5]

[a] Universitas Ahmad Dahlan, Indonesia

[1] almuzhidul1800015044@math.uad.ac.id*; [2] aris.thobi@math.uad.ac.id ; [3]salma1800015090@math.uad.ac.id; [4] sugiyarto@math.uad.ac.id; [5] lanova1800015088@webmail.uad.ac.id
*Correspondent Author

KEYWORDS

KNN
Analisis Sentimen
Ulasan

ABSTRAK

K-Nearest Neighbor (KNN) merupakan algoritma yang biasa digunakan untuk klasifikasi. Penelitian ini menggunakan ulasan aplikasi Maxim di Google Play Store. Pengguna yang sudah mengunduh aplikasi Maxim berhak memberikan ulasan di Google Play Store guna berbagi informasi untuk pengguna lain. Implementasi K-Nearest Neighbor (KNN) terhadap Sentiment Analysis ulasan aplikasi Maxim dapat digunakan untuk menentukan kelas ulasan bernilai positif, neutral, atau negatif. Peneliti melakukan perbandingan 5 jarak yang berbeda untuk metode KNN yaitu jarak Euclidean, Manhattan, Minkowski, Chebyshev dan Canberra. Pengujian yang telah dilakukan memberikan hasil akurasi pada klasifikasi KNN dengan jarak yang berbeda, memberikan hasil akurasi yang berbeda-beda, yaitu jarak Euclidean = 84%, jarak Manhattan = 79%, jarak Minkowski 84%, jarak Chebyshev = 75% dan jarak Canberra = 44%.

Forecasting Rainfall in Bandung City Using Singular Spectrum Analysis

KEYWORDS

KNN
Sentiment Analysis
Review

K-Nearest Neighbor (KNN) is an algorithm commonly used for classification. This study uses a review of the Maxim application on the Google Play Store. Users who have downloaded the Maxim application have the right to leave reviews on the Google Play Store to share information with other users. The implementation of K-Nearest Neighbor (KNN) on the Sentiment Analysis of Maxim application reviews can be used to determine the class of reviews that are positive, neutral, or negative. Researchers compared 5 different distances for the KNN method, namely the Euclidean, Manhattan, Minkowski, Chebyshev and Canberra distances. The tests that have been carried out provide accuracy results in the KNN classification with different distances, giving different accuracy results, namely Euclidean distance = 84%, Manhattan distance = 79%, Minkowski distance 84%, Chebyshev distance = 75% and Canberra distance = 44%.

 ISSN 2774-3241

Pendahuluan

Meningkatnya penggunaan media elektronik bersamaan dengan perkembangan teknologi di Indonesia yang sangat membantu aktivitas manusia, salah satunya pada sektor transportasi [1]. Transportasi di Indonesia yang banyak diperbincangkan adalah transportasi online, karena sangat efisien dan bisa mengurangi tingginya angka kemacetan di Indonesia. Dahulu apabila ingin menggunakan transportasi taksi atau ojek, harus ke pangkalan ojek terlebih dahulu [2]. Sekarang dapat dengan mudah memesan transportasi online menggunakan smartphone maka transportasi online akan tiba menghampiri dan siap mengantar sampai tujuan tanpa membicarakan kesepakatan harga, salah satu penyedia transportasi online adalah Maxim [3].

Maxim merupakan perusahaan transportasi online baru di Indonesia. Maxim dirintis pada tahun 2003 di Rusia. Maxim sendiri pertama kali beroperasi di Indonesia pada tahun 2018, Maxim saat ini juga sudah mulai tersebar di berbagai kota besar di Indonesia, seperti; Jakarta, Pekanbaru, Batam, Bandar Lampung, Yogyakarta, Solo, Samarinda, Banjarmasin, Pontianak, dan Bali [4].

Pada Google Playstore setiap pengguna yang mengunduh aplikasi berhak memberikan informasi berupa rating dan ulasan terhadap aplikasi yang diunduh. Informasi dan ulasan suatu produk disimpan dalam bentuk teks, maka solusi yang dapat digunakan dalam pengambilan informasi yang berbentuk teks adalah menggunakan *text mining* [5]. Dalam mengkategorikan teks pada proses pengambilan informasi *text mining* salah satunya dengan *sentiment analysis* yang digunakan untuk klasifikasi sebuah informasi suatu teks bahasa kedalam kategori positif, neutral atau negatif [6].

Pada penelitian sebelumnya yang dilakukan oleh [1] berjudul 'Analisis Sentimen Aplikasi Gojek Menggunakan Support Vector Machine Dan K Nearest Neighbor' diperoleh hasil akurasi sebesar 82,14% saat melakukan klasifikasi dengan K- Nearest Neighbor.

Algoritma klasifikasi K-Nearest Neighbor dilatih untuk melihat pola ulasan dari pengguna aplikasi, selanjutnya dari hasil kelas ulasan maka akan diketahui sentimen dari pengguna, yang kemudian sistem ini lebih lanjut dapat digunakan untuk membalas ulasan menggunakan *auto answer*. Proses penelitian dilakukan dengan melakukan scraping data dengan menggunakan bahasa pemrograman python lalu data disimpan dengan format csv yang berisi kolom username, ulasan, waktu dan rating, sehingga membentuk sebuah dataset yang dapat digunakan.

Data yang akan digunakan untuk simulasi perhitungan adalah 1.000 data ulasan aplikasi Maxim pada Google Play Store. Penelitian ini akan dilakukan perbandingan 5 jarak pada metode KNN untuk mendapatkan hasil terbaik pada proses klasifikasi. Jarak yang akan digunakan pada proses klasifikasi KNN diantara yaitu jarak Euclidean, jarak Manhattan, jarak Minkowski, jarak Chebyshev dan jarak Canberra.

Metode

Natural Language Processing adalah suatu bidang penelitian moderen dalam ilmu komputer dan kecerdasan buatan (AI) yang berkaitan dengan pemrosesan bahasa alami seperti bahasa Inggris atau Mandarin. Pemrosesan ini umumnya melibatkan penerjemahan bahasa alami ke dalam data (angka) yang dapat digunakan oleh komputer untuk mempelajari kata. Dan pemahaman tentang kata ini terkadang

 ISSN 2774-3241

digunakan untuk menghasilkan teks bahasa alami yang mencerminkan pemahaman tersebut [7], [8].

Text Mining

Text mining juga dikenal sebagai penemuan pengetahuan atau informasi dalam database tekstual atau penambangan data teks, di mana pengetahuan baru yang menarik dibuat, didefinisikan sebagai proses mengekstraksi yang sebelumnya tidak diketahui dari kumpulan data teks atau korpus yang masif dan tidak terstruktur. Text mining diyakini memiliki nilai komersial yang lebih tinggi daripada cabang data mining lainnya, karena 80% informasi perusahaan terkandung dalam dokumen teks. Namun, *text mining* lebih kompleks karena data teks tidak terstruktur. *Text mining* adalah area penelitian yang komprehensif, yang melibatkan bidang kecerdasan buatan, pembelajaran mesin, statistik matematika, sistem basis data, dan sebagainya [9].

Pre-processing

Pre-processing (Pra-pemrosesan) merupakan tahapan yang dilakukan sebelum pengolahan data. *pre-processing* kali ini dilakukan dengan tujuan untuk mengubah teks pada dokumen ke dalam bentuk teks lain supaya lebih mudah diolah pada pengolahan data, sehingga dapat memberikan hasil yang lebih optimal [10]. *Preprocessing* memiliki 4 tahapan, diantaranya:

Casefolding

Fitur *Casefolding* adalah proses yang digunakan untuk mengubah huruf kapital menjadi huruf kecil [11]. Contohnya pada kalimat 'Saya sedang makan' menjadi 'saya sedang makan'.

Tokenization

Tokenization adalah proses yang dilakukan memecahkan urutan kalimat menjadi potongan-potongan kata untuk proses pengambilan kata pada kalimat, masing-masing kata disebut dengan *token* [12]. Contoh pada kalimat 'Budi pergi ke pasar' memiliki 4 *token* yaitu: Budi, pergi, ke, sekolah.

Stopwording

Stopwording merupakan proses penghilangan kata, karena *stopword* adalah kata-kata yang dianggap tidak ada artinya [13]. Kata yang dianggap tidak ada arti biasanya adalah kata hubung dan kata sambung, contohnya: di, dan, dari, ke, ini, pada, dll.

Stemming

Stemming digunakan untuk mengidentifikasi dan menghapus kata imbuhan atau bisa disebut juga proses pengambilan kata dasar dari kata yang berimbuhan [14]. Contohnya kata 'mengucap' menjadi 'ucap', kata 'terjatuh' menjadi 'jatuh'.

BoW dan *TF-IDF*

Bag of Words (BoW) yaitu konsep dari analisis teks yang merepresentasikan sebuah dokumen sebagai kantung informasi penting tanpa harus mengurutkan setiap katanya.

TF-IDF yaitu metode pembobotan kata dalam proses pengolahan teks. Metode TF-IDF terkenal sangat efisien, mudah , dan memiliki hasil yang akurat, dimana metode ini akan menghitung nilai *Term of Frequency* (TF) dan *Inverse Document Frequency* (IDF) pada setiap kata yang ada di tiap dokumen dalam korpus.

$$IDF(w) = log(NDF(w))$$
$$TFIDF(w,d) = TF(w,d).IDF(W)$$

Keterangan :

$TFIDF(w,d)$ = bobot suatu kata dalam keseluruhan dokumen
w = suatu kata (word)
d = suatu dokumen
$TF(w,d)$ = frekuensi kemunculan sebuah kata w dalam dokumen
$IDF(w)$ = invers DF dari kata w
pN = jumlah keseluruhan dokumen
$DF(w)$ = jumlah dokumen yang mengandung kata w

K-Nearest Neighbors (KNN) merupakan salah satu metode pengambilan keputusan menggunakan pembelajaran terkontrol dimana hasil dari data *input* yang baru diklasifikasikan berdasarkan jarak terdekat dengan kumpulan data sebelumnya [15]. KNN merupakan metode yang menggunakan algoritma supervised dengan hasil dari query instance yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada *K-Nearest Neighbors.* Menurut [16] *Nearest Neighbors* yaitu suatu pendekatan untuk mencari kelas data dengan menghitung kedekatan antara data baru dengan data lama, yaitu dengan berdasarkan pada pencocokan bobot dari sejumlah fitur yang ada. Teknik metode ini adalah algoritma *machine learning* yang dianggap mudah untuk diimplementasikan. Metode ini tergolong dalam kelompok *instance-based learning*, yang merupakan salah satu teknik *lazy learning*, dengan mencari kelompok k objek dalam data latih yang paling dekat (mirip) dengan objek pada data uji [17]. Model klasifikasi metode ini, prediksinya murni berdasarkan nilai dataset. K mewakili jumlah nilai tetangga terdekatnya sebagai keputusan untuk mengklasifikasikan kumpulan data yang diberikan. Tujuannya yaitu untuk mengklasifikasikan data baru berdasarkan variabel(*feature)* dan sampel data latih. Prinsip umumnya adalah menemukan k data uji untuk menentukan kelas berdasarkan ukuran jarak.

Langkah-langkah untuk menghitung metode *K-Nearest Neighbor* [18] antara lain:

1.) Menentukan parameter K (jumlah tetangga paling dekat).

2.) Menghitung kuadrat jarak (*query instance*) masing-masing objek terhadap data sampel yang diberikan menggunakan persamaan rumus jarak Euclidean. Untuk menghitung jarak antara dua titik x dan y digunakan 5 metode perhitungan jarak, yaitu Euclidean, Manhattan, Minkowski, Chebyshev, dan Canberra.

- Euclidean

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

Keterangan :
d_{ij} = jarak antara objek i dengan j
x_{ik} = nilai objek i pada variabel ke-k
x_{jk} = nilai objek j pada variabel ke-k
p = banyaknya variabel yang diamati

- Manhattan

$$d(x,y) = \sum_{k=1}^{n} |x_i - y_i|$$

dimana,
d = jarak antara objek i dengan j

 ISSN 2774-3241

x = nilai objek i pada variabel ke-k
y = nilai objek j pada variabel ke-k

- Minkowski

$$d(x,y) = \left(\sum_{k=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

dimana,
d = jarak antara objek x dengan y
x = data pusat klaster
y = data pada atribut
i = setiap data
n = jumlah data,
x_i = data pada pusat klaster ke i
y_i = dat pada setiap data ke i
p = power

- Chebyshev

$$d_{ij} = \max_{k}(|x_{ik} - y_{jk}|)$$

dimana,
d = jarak antara i dan j
i = pusat data cluster
j = data pada atribut
k = simbol setiap data

- Canberra

$$d_{ij} = \sum_{k=1}^{n} \frac{|x_{ik} - y_{jk}|}{|x_{ik}| + |y_{jk}|}$$

dimana,
d = jarak antara i dan j
i = pusat data cluster
j = data pada atribut
k = simbol setiap data
n = jumlah data

Untuk menentukan nilai k yang tepat, maka diperlukan akurasi yang tinggi dalam proses kategorisasi dokumen uji. ketika nilai k semakin tinggi, hasil kategori tidak terpengaruhi pada kategori yang memiliki dokumen latih yang lebih besar, karena perbedaan nilai k yang dimiliki setiap kategori dapat disesuaikan dengan besar - kecilnya dari banyaknya dokumen latih yang lebih besar.

3.) Kemudian mengurutkan objek-objek tersebut ke dalam kelompok yang mempunyai jarak terkecil.

4.) Mengumpulkan kategori y (klasifikasi *K-Nearest Neighbor*).

5.) Menggunakan kategori *K-Nearest Neighbor* yang paling mayoritas untuk mendapatkan prediksi nilai *query instance* yang telah dihitung.

 ISSN 2774-3241

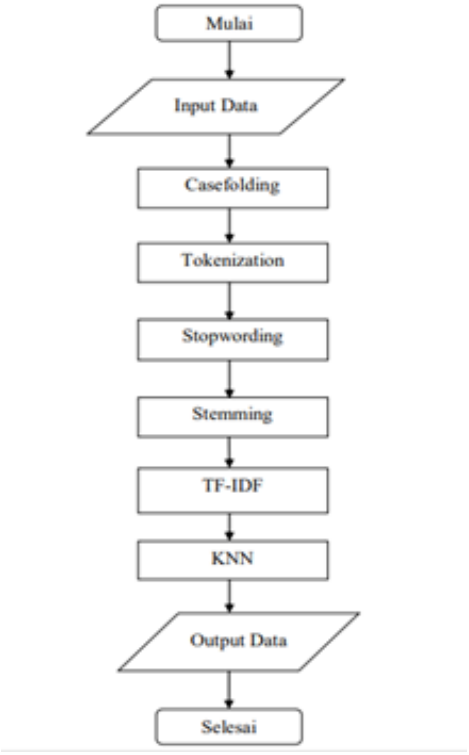
Hasil dan Pembahasan

1. Pengumpulan Data

Pengumpulan data dilakukan dengan mengumpulkan data ulasan aplikasi Maxim dengan cara scrapping data menggunakan bahasa pemrograman python. Mengulas aplikasi merupakan cara yang efektif guna berbagi masukan yang bermanfaat dan dapat dipercaya untuk membantu pengguna maxim lain mempertimbangkan pengunduhan aplikasi. Pada tahap ini, data terdiri dari 1.000 ulasan aplikasi Maxim yang akan digunakan untuk proses selanjutnya. Data yang didapatkan disimpan dalam sebuah file teks dengan format csv yang berisi kolom username, rating, waktu dan ulasan.

2. Pembangunan Sistem

Sistem dirancang menggunakan bahasa pemrograman python. Tahap pertama dalam pembangunan sistem pada pemrograman python adalah diawali dengan melakukan input data yang akan digunakan yaitu data ulasan aplikasi Maxim. Tahap selanjutnya yaitu melakukan *pre-processing*, pada *pre-processing* terdapat 4 tahap, yaitu: *casefolding, tokenization, stopwording, stemming*. Hasil yang didapatkan dari tahap *pre-processing* akan digunakan pada proses selanjutnya yaitu pada tahap pembobotan kata menggunakan perhitungan TI-IDF. Tahap terakhir hasil dari pembobotan kata yang menggunakan perhitungan TI-IDF akan diklasifikasi menggunakan KNN, tahap klasifikasi KNN peneliti menggunakan perhitungan dengan membandingkan 5 metode perhitungan jarak, yaitu Euclidean, Manhattan, Minkowski, Chebyshev, dan Canberra. Diagram alir mengenai cara kerja pembangunan sistem dapat dilihat pada gambar 1.



Gambar 1. Diagram alir mengenai cara kerja pembangunan sistem

 ISSN 2774-3241

3. Perhitungan

Sistem dirancang menggunakan Pada saat melakukan eksperimen menggunakan metode K-Nearest Neighbor, peneliti menggunakan data sebanyak 1.000 ulasan yang paling relevan. Tahap dasar yang dilakukan pada proses klasifikasi yaitu:

a. *Preprocessing*

Table.1 CASEFOLDING

Teks Sebelum Proses Casefolding	***Teks Setelah Proses Casefolding***
Driver Ramah Dan Baik	driver ramah dan baik
Komen buruk sekali	komen buruk sekali

Table.2 TOKENIZING

Teks Sebelum Proses Tokenization	***Teks Setelah Proses Tokenization***
driver ramah dan baik	driver, ramah, dan, baik
komen buruk sekali	komen, buruk, sekali

Table.3 STOPWORDING

Teks Sebelum Proses Stopwording	***Teks Setelah Proses Stopwording***
driver, ramah, dan, baik	driver, ramah, baik
komen, buruk, sekali	komen, buruk, sekali

Table.4 *STEMMING*

Teks Sebelum Proses Stemming	***Teks Setelah Proses Stemming***
driver, ramah, baik	‘driver’, ‘ramah’, ‘baik’
komen, buruk, sekali	‘komen’, ‘buruk’, ‘sekali’

Table.5 DATA CONTOH SETELAH *PREPROCESSING*

Data	***Sebelum Proses Preprocessing***	***Setelah Proses Preprocessing***	***Kelas***
U1	kurang ramah	[ramah]	negatif
U2	thankyou bosque👍	[thankyou, bosque]	positif
U3	bang kurirnya baik hati ramah dan sabar, 😁👍 terimakasih bang kurir, maaf sdh merepotkn	[bang, kurir, hati, ramah, sabar, terimakasih, bang, kurir, maaf, merepotkan]	positif

 ISSN 2774-3241

U4	Komen ya pd buruk sekli ya	[komen, buruk, sekli]	negatif
U5	sudah bagus, ditingkatkan lagi akurasi petanya. bravo 👍	[bagus, tingkat, akurasi, bravo]	positif
U6	Akurasi kurir buruk	[akurasi, kurir, buruk]	?

b. Menghitung jumlah frekuensi tiap kata pada tiap dokumen (TF)

Table.6 HASIL PEMBOBOTAN KATA DATA CONTOH

Kata	***D6***	***D1***	***D2***	***D3***	***D4***	***D5***	***DF***
ramah	0	1	0	1	0	0	2
thankyou	0	0	1	0	0	0	1
bosque	0	0	1	0	0	0	1
bang	0	0	0	2	0	0	2
kurir	1	0	0	2	0	0	3
hati	0	0	0	1	0	0	1
sabar	0	0	0	1	0	0	1
terimakasih	0	0	0	1	0	0	1
maaf	0	0	0	1	0	0	1
merepotkn	0	0	0	1	0	0	1
komen	0	0	0	0	1	0	1
buruk	1	0	0	0	1	0	2
sekli	0	0	0	0	1	0	1
bagus	0	0	0	0	0	1	1
tingkat	0	0	0	0	0	1	1
akurasi	1	0	0	0	0	1	2
bravo	0	0	0	0	0	1	1

c. Menghitung IDF menggunakan persamaan $IDF(w) = log(NDF(w))$

Table.7 HASIL PERHITUNGAN IDF DATA CONTOH

Kata	***IDF***
ramah	$^{10}\log(62)=0.48$
thankyou	$^{10}\log(61)=0.78$
bosque	$^{10}\log(61)=0.78$

 ISSN 2774-3241

bang	^{10}log(62)=0.48
kurir	^{10}log(63)=0.3
hati	^{10}log(61)=0.78
sabar	^{10}log(61)=0.78
terimakasih	^{10}log(61)=0.78
maaf	^{10}log(61)=0.78
merepotkn	^{10}log(61)=0.78
komen	^{10}log(61)=0.78
buruk	^{10}log(62)=0.48
sekli	^{10}log(61)=0.78
bagus	^{10}log(61)=0.78
tingkat	^{10}log(61)=0.78
akurasi	^{10}log(62)=0.48
bravo	^{10}log(61)=0.78

d. Menghitung TF-IDF menggunakan persamaaan $TFIDF(w,d) = TF(w,d).IDF(W)$

Table.8 HASIL PERHITUNGAN TF-IDF DATA CONTOH

Kata	***U6***	***U1***	***U2***	***U3***	***U4***	***U5***
ramah	0	0.48	0	0.48	0	0
thankyou	0	0	0.78	0	0	0
bosque	0	0	0.78	0	0	0
bang	0	0	0	0.9	0	0
kurir	0.3	0	0	0.9	0	0
hati	0	0	0	0.78	0	0
sabar	0	0	0	0.78	0	0
terimakasih	0	0	0	0.78	0	0
maaf	0	0	0	0.78	0	0
merepotkn	0	0	0	0.78	0	0
komen	0	0	0	0	0.78	0
buruk	0.48	0	0	0	0.48	0
sekli	0	0	0	0	0.78	0
bagus	0	0	0	0	0	0.78
tingkat	0	0	0	0	0	0.78
akurasi	0.48	0	0	0	0	0.48
bravo	0	0	0	0	0	0.78

e. Perhitungan jarak Euclidean

$$d(U1,U6) = \sqrt{\begin{matrix}(0-0.48)2+(0-0)2+(0-0)2+(0-0)2+(0,3-0)2+(0-0)2+\\(0-0)2+(0-0)2+(0-0)2+(0-0)2+(0-0)2+(0.48-0)2+\\(0-0)2+(0-0)2+(0-0)2+(0.48-0)2+(0-0)2\end{matrix}}$$

$$= \sqrt{0{,}7812}$$
$$= 0.8838\ (urutan\ 1)$$

$$d(U2,U6) = \sqrt{\begin{matrix}(0-0.48)2+(0.78-0)2+(0.78-0)2+(0-0)2+(0.3-0)2+\\ +(0-0)2+(0-0)2+(0-0)2+(0-0)2+(0-0)2+(0-0)2+\\ (0.48-0)2+(0-0)2+(0-0)2+(0-0)2+(0-0.48)2+(0-0)2\end{matrix}}$$
$$= \sqrt{1.7676}$$
$$= 1.3295\ (urutan\ 2)$$

$$d(U3,U6) = \sqrt{\begin{matrix}(0-0{,}48)2+(0-0)2+(0-0)2+(0-0{,}9)2+(0{,}3-0{,}9)2+(0-0{,}78)2\\ +(0-0{,}78)2+(0-0{,}78)2+(0-0{,}78)2+(0-0{,}78)2+(0-0)2+\\ (0{,}48-0)2+(0-0)2+(0-0)2+(0-0)2+(0{,}48-0)2+(0-0)2\end{matrix}}$$
$$= \sqrt{4.69584}$$
$$= 2.166\ (urutan\ 5)$$

$$d(U4,U6) = \sqrt{\begin{matrix}(0-0)2+(0-0)2+(0-0)2+(0-0)2+(0.3-0)2+(0-0)2+\\ (0-0)2+(0-0)2+(0-0)2+(0-0)2+(0-0.78)2+(0-0.48)2+\\ (0-0.78)2+(0-0)2+(0-0)2+(0-0)2+(0-0)2\end{matrix}}$$
$$= \sqrt{1{,}7868}$$
$$= 1.336\ (urutan\ 3)$$

$$d(U5,U6) = \sqrt{\begin{matrix}(0-0)2+(0-0)2+(0-0)2+(0-0)2+(0{,}3-0)2+(0-0)2+\\ (0-0)2+(0-0)2+(0-0)2+(0-0)2+(0-0)2+(0-0)2+\\ (0.48-0)2+(0-0)2+(0-0.78)2+(0-0.78)2+(0-0.48)2\\ +(0-0.78)2\end{matrix}}$$
$$= \sqrt{2.376}$$
$$= 1.541\ (urutan\ 4)$$

Manhattan

$$d(U1,U6) = |0{,}48-0|+|0-0|+|0-0|+|0-0|+|0-0.3|+|0-0|+|0-0|+|0-0|+|0-0|+|0-0|+|0-0|+|0-0|+|0-0.48|+|0-0|+|0-0|+|0-0|+|0-0.48|+|0-0|$$
$$= 1.74\ (urutan\ 2)$$

$$d(U2,U6) = |0-0|+|0.78-0|+|0.78-0|+|0-0|+|0-0.3|+|0-0|+|0-0|+|0-0|+|0-0|+|0-0|+|0-0|+|0-0|+|0-0.48|+|0-0|+|0-0|+|0-0|+|0-0.48|+|0-0|$$
$$= 1.56\ (urutan\ 1)$$

$$d(U3,U6) = |0{,}48-0|+|0-0|+|0-0|+|0.9-0|+|0.9-0.3|+|0.78-0|+|0.78-0|+|0.78-0|+|0.78-0|+|0-0|+|0-0.48|+|0-0|+|0-0|+|0-0|+|0-0.48|+|0-0|$$
$$= 6.84\ (urutan\ 5)$$

$$d(U4,U6) = |0-0|+|0-0|+|0-0|+|0-0|+|0-0.3|+|0-0|+|0-0|+|0-0|+|0-0|+|0-0|+|0.78-0|+|0.48-0.48|+|0.78-0|+|0-0|+|0-0|+|0-0.48|+|0-0|$$
$$= 2.82\ (urutan\ 4)$$

$$d(U5,U6) = |0-0|+|0-0|+|0-0|+|0-0|+|0-0.3|+|0-0|+|0-0|+|0-0|+|0-0|+|0-0|+|0-0|+|0-0|+|0-0|+|0.78-0|+|0.78-0|+|0.48-0.48|+|0.78-0|$$
$$= 2.34\ (urutan\ 3)$$

Minkowski

$$d(U1,U6) = (|0.48-0|^4+|0-0|^4+|0-0|^4+|0-0|^4+|0-0.3|^4+|0-0|^4+|0-0|^4+|0-0|^4+|0-0|^4+|0-0|^4+|0-0|^4+|0-0|^4+|0-0|^4+|0-0.48|^4+|0-0|^4+|0-0|^4+|0-0|^4+|0-0.48|^4+|0-0|^4)^{1/4}$$
$$= 0.6396\ (urutan\ 1)$$

 ISSN 2774-3241

$$d(U2,U6) = (|0-0|^4 + |0.78-0|^4 + |0.78-0|^4 + |0-0|^4 + |0-0.3|^4 + |0-0|^4 + |0-0|^4 + |0-0|^4 + |0-0|^4 + |0-0|^4 + |0-0|^4 + |0-0.48|^4 + |0-0|^4 + |0-0|^4 + |0-0|^4 + |0-0.48|^4 + |0-0|^4)^{1/4}$$
$$= 0.9615\ (urutan\ 3)$$
$$d(U3,U6) = (|0.48-0|^4 + |0-0|^4 + |0-0|^4 + |0.9-0|^4 + |0.9-0.3|^4 + |0.78-0|^4 + |0.78-0|^4 + |0.78-0|^4 + |0.78-0|^4 + |0.78-0|^4 + |0-0|^4 + |0-0.48|^4 + |0-0|^4 + |0-0|^4 + |0-0|^4 + |0-0|^4 + |0-0.48|^4 + |0-0|^4)^{1/4}$$
$$= 1.2930\ (urutan\ 5)$$
$$d(U4,U6) = (|0-0|^4 + |0-0|^4 + |0-0|^4 + |0-0|^4 + |0-0.3|^4 + |0-0|^4 + |0-0|^4 + |0-0|^4 + |0-0|^4 + |0-0|^4 + |0.78-0|^4 + |0.48-0.48|^4 + |0.78-0|^4 + |0-0|^4 + |0-0|^4 + |0-0.48|^4 + |0-0|^4)^{1/4}$$
$$= 0.9462\ (urutan\ 2)$$
$$d(U5,U6) = (|0-0|^4 + |0-0|^4 + |0-0|^4 + |0-0|^4 + |0-0.3|^4 + |0-0|^4 + |0-0|^4 + |0-0|^4 + |0-0|^4 + |0-0|^4 + |0-0|^4 + |0-0.48|^4 + |0-0|^4 + |0.78-0|^4 + |0.78-0|^4 + |0.48-0.48|^4 + |0.78-0|^4)^{1/4}$$
$$= 1.0404\ (urutan\ 4)$$

Chebyshev

$$d(U1,U6) = \max_k \begin{pmatrix} |0.48-0| + |0-0| + |0-0| + |0-0| + |0-0.3| + |0-0| + |0-0| + \\ |0-0| + |0-0| + |0-0| + |0-0| + |0-0.48| + |0-0| + \\ |0-0| + |0-0| + |0-0.48| + |0-0| \end{pmatrix}$$
$$= \max_k(1.74)\ (urutan\ 1)$$
$$d(U2,U6) = \max_k \begin{pmatrix} |0-0| + |0.78-0| + |0.78-0| + |0-0| + |0-0.3| + |0-0| + \\ |0-0| + |0-0| + |0-0| + |0-0| + |0-0| + |0-0.48| + |0-0| + \\ |0-0| + |0-0| + |0-0.48| + |0-0| \end{pmatrix}$$
$$= \max_k(12.82)\ (urutan\ 5)$$
$$d(U3,U6) = \max_k \begin{pmatrix} |0.48-0| + |0-0| + |0-0| + |0.9-0| + |0.9-0.3| + |0.78-0| + \\ |0.78-0| + |0.78-0| + |0.78-0| + |0.78-0| + |0-0| + |0-0.48| \\ +|0-0| + |0-0| + |0-0| + |0-0.48| + |0-0| \end{pmatrix}$$
$$= \max_k(6.48)\ (urutan\ 4)$$
$$d(U4,U6) = \max_k \begin{pmatrix} |0-0| + |0-0| + |0-0| + |0-0| + |0-0.3| + |0-0| + \\ |0-0| + |0-0| + |0-0| + |0-0| + |0.78-0| + |0.48-0.48| \\ +|0.78-0| + |0-0| + |0-0| + |0-0.48| + |0-0| \end{pmatrix}$$
$$= \max_k(2.34)\ (urutan\ 2)$$
$$d(U5,U6) = \max_k \begin{pmatrix} |0-0| + |0-0| + |0-0| + |0-0| + |0-0.3| + |0-0| + \\ |0-0| + |0-0| + |0-0| + |0-0| + |0-0| + |0-0.48| \\ +|0-0| + |0.78-0| + |0.78-0| + |0.48-0.48| + |0.78-0| \end{pmatrix}$$
$$= \max_k(3.12)\ (urutan\ 3)$$

Canberra

$$d(U1,U6) = \frac{|0.48-0|}{|0.48|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0.3|}{|0|+|0.3|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0.48|}{|0|+|0.48|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0.48|}{|0|+|0.48|} + \frac{|0-0|}{|0|+|0|}$$
$$= 4\ (urutan\ 1)$$

 ISSN 2774-3241

$$d(U2,U6) = \frac{|0-0|}{|0|+|0|} + \frac{|0.78-0|}{|0.78|+|0|} + \frac{|0.78-0|}{|0.78|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0.3|}{|0|+|0.3|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0.48|}{|0|+|0.48|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0.48|}{|0|+|0.48|} + \frac{|0-0|}{|0|+|0|}$$

$$= 5\ (urutan\ 3)$$

$$d(U3,U6) = \frac{|0.48-0|}{|0.48|+|0|} + \frac{|0.78-0|}{|0.78|+|0|} + \frac{|0.78-0|}{|0.78|+|0|} + \frac{|0.9-0|}{|0.9|+|0|} + \frac{|0.9-0.3|}{|0.9|+|0.3|} + \frac{|0.78-0|}{|0.78|+|0|} + \frac{|0.78-0|}{|0.78|+|0|} + \frac{|0.78-0|}{|0.78|+|0|} + \frac{|0.78-0|}{|0.78|+|0|} + \frac{|0.78-0|}{|0.78|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0.48|}{|0|+|0.48|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0.48|}{|0|+|0.48|} + \frac{|0-0|}{|0|+|0|}$$

$$= 16\ (urutan\ 4)$$

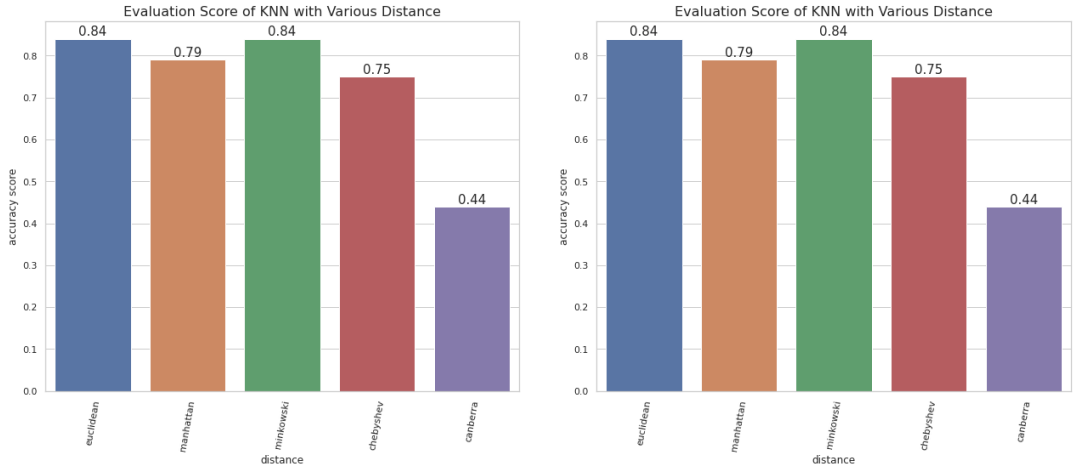
$$d(U4,U6) = \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0.3|}{|0|+|0.3|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0.78-0|}{|0.78|+|0|} + \frac{|0.78-0.48|}{|0.78|+|0.48|} + \frac{|0.78-0|}{|0.78|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0.48|}{|0|+|0.48|} + \frac{|0-0|}{|0|+|0|}$$

$$= 4.238\ (urutan\ 2)$$

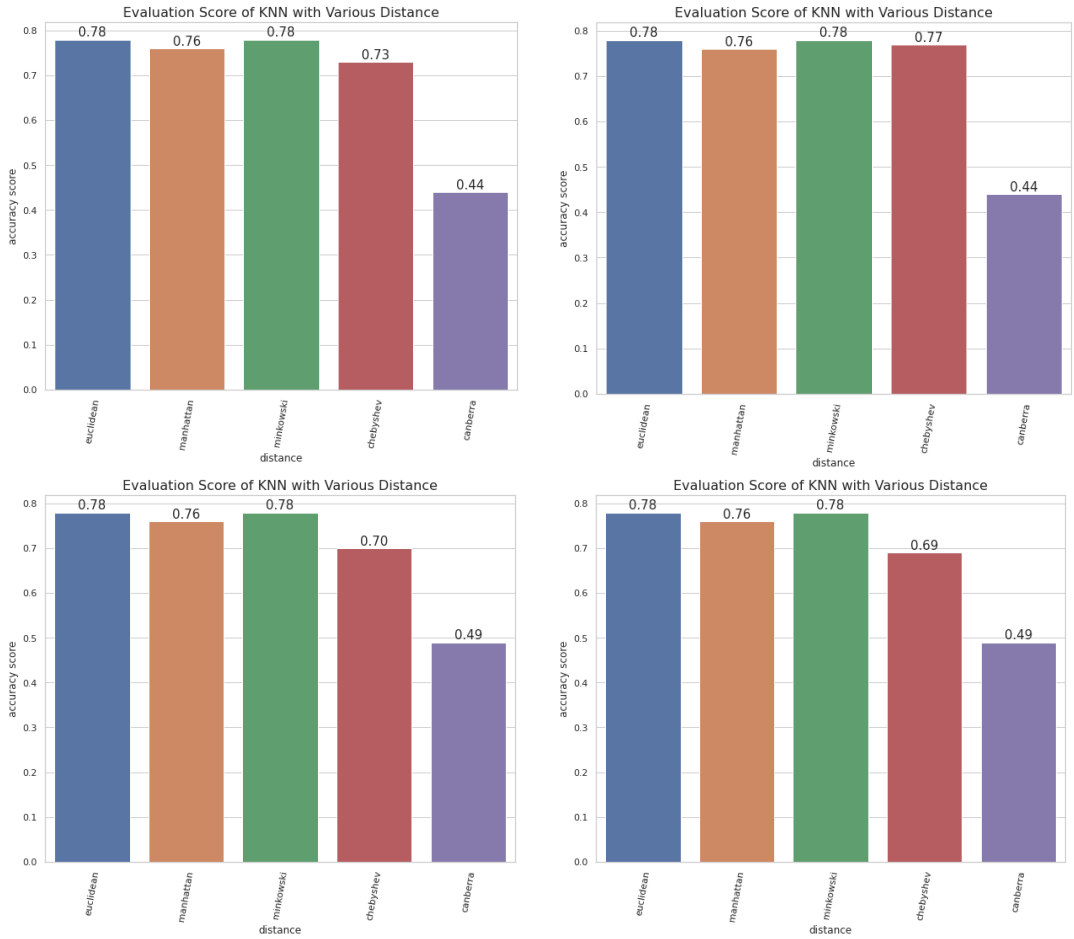
$$d(U5,U6) = \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0.3|}{|0|+|0.3|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0|}{|0|+|0|} + \frac{|0-0.48|}{|0|+|0.48|} + \frac{|0-0|}{|0|+|0|} + \frac{|0.78-0|}{|0.78|+|0|} + \frac{|0.78-0|}{|0.78|+|0|} + \frac{|0.48-0.48|}{|0.48|+|0.48|} + \frac{|0.48-0|}{|0.48|+|0|}$$

$$= 5\ (urutan\ 3)$$

Dengan menggunakan 1.000 ulasan didapatkan perbandingan akurasi masing-masing jarak, untuk K yang bervariasi seperti pada gambar 2 dibawah,



 ISSN 2774-3241



Gambar 2. Grafik akurasi K=1 (*kiri atas*),K=2 (*kanan atas*),K=3 (*kiri tengah*),K=4 (*kanan tengah*),K=5 (*kiri bawah*), dan K=6 (*kanan bawah*),

Berdasarkan perubahan nilai K yang sudah dilakukan, nilai akurasi tertinggi terdapat pada K=1 dan K=2, dengan akurasi yang berbeda untuk masing-masing metode jarak yang digunakan, yaitu akurasi terbaik didapatkan dengan menggunakan jarak Euclidean dan Minkowski.

Simpulan

Berdasarkan hasil implementasi dan pengujian yang telah dilakukan pada metode KNN dengan menggunakan perbandingan jarak, dapat disimpulkan bahwa: Perbandingan jarak pada klasifikasi KNN menghasilkan nilai akurasi masing-masing sebesar 84% untuk jarak Euclidean, 79% untuk jarak Manhattan, 84% untuk jarak Minkowski, 75% untuk jarak Chebyshev dan 44% untuk jarak Canberra. Nilai akurasi tertinggi didapatkan saat menggunakan jarak Euclidean dan jarak Minkowski pada proses klasifikasi KNN. Untuk penelitian kedepannya diharapkan dapat menerapkan tahap preprocessing yang lebih baik, sehingga dapat dapat memilih kata dasar berdasarkan konteks ulasan, agar dapat memberikan akurasi yang lebih baik pada proses klasifikasi.

 ISSN 2774-3241

Daftar Pustaka

[1] M. N. Muttaqin and I. Kharisudin, "UNNES Journal of Mathematics," *UNNES J. Math.*, vol. 1, no. 2252, pp. 125–130, 2015.

[2] P. B. Mahargiono and K. E. Cahyono, "Kontroversi Transportasi Online Sebagai Dasar Pembenahan Fasiltas Layanan Penumpang Bagi Pelaku Bisnis Transportasi Di Surabaya," *Pros. Semin. Nas. Multi Disiplin Ilmu*, vol. 3, no. Sendi_U 3, pp. 663–668, 2017.

[3] F. Al Rasyid, "Resahkan' Kompetitor, Maxim Ternyata Bukan Perusahaan Ojek Online," [Online]. Available: https://id.rbth.com/economics/81796-maxim-ojol-asalrusia-wyx.

[4] A. Indriani, "Mengenal Ojol Maxim, Penantang Gojek-Grab," *DetikFinance*, 2019.

[5] N. Wayan and S. Saraswati, "Naïve bayes classifier dan support vector machines untuk sentiment analysis," *Semin. Nas. Sist. Inf. Indones.*, pp. 2–4, 2013.

[6] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining Text Data*, 2012, pp. 415–463.

[7] M. Mukaromah, "Penerapan Metode Fuzzy Sugeno Untuk Menentukan Jalur Terbaik Menuju Lokasi Wisata Di Surabaya," *J. Mat. Sains dan Teknol.*, vol. 20, no. 2, pp. 95–101, 2019, doi: 10.33830/jmst.v20i2.187.2019.

[8] Z. Putri, Sugiyarto, and Salafudin, "Desimal : Jurnal Matematika," *Desimal J. Mat.*, vol. 4, no. 1, pp. 13–20, 2021, doi: 10.24042/djm.

[9] Y. Zhang, M. Chen, and L. Liu, "A review on text mining," *Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS*, vol. 2015-November, pp. 681–685, 2015, doi: 10.1109/ICSESS.2015.7339149.

[10] K. Nugraha and D. Sebastian, "Chatbot Layanan Akademik Menggunakan K-Nearest Neighbor," *J. Sains dan Inform.*, vol. 7, no. 1, pp. 11–19, 2021, [Online]. Available: https://dev.jsi.politala.ac.id/index.php/JSI/article/view/285.

[11] H. . Putranoto, T. Rizaldi, W. K. Dewanto, and W. Pebrianto, "DESIGNING A PYTHON BASED TEXT PRE-PROCESSING APPLICATION FOR TEXT CLASSIFICATION," pp. 187–194.

[12] A. Jettakul, C. Thamjarat, K. Liaowongphuthorn, C. Udomcharoenchaikit, P. Vateekul, and P. Boonkwan, "A Comparative Study on Various Deep Learning Techniques for Thai NLP Lexical and Syntactic Tasks on Noisy Data," *Proceeding 2018 15th Int. Jt. Conf. Comput. Sci. Softw. Eng. JCSSE 2018*, pp. 1–6, 2018, doi: 10.1109/JCSSE.2018.8457368.

[13] D. J. Ladani and N. P. Desai, "Stopword Identification and Removal Techniques on TC and IR applications: A Survey," *2020 6th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2020*, pp. 466–472, 2020, doi: 10.1109/ICACCS48705.2020.9074166.

[14] H. M. Tahir, A. M. Said, N. H. Osman, N. H. Zakaria, P. N. A. M. Sabri, and N. Katuk, "Improving K-Means Clustering using discretization technique in Network Intrusion Detection System," *2016 3rd Int. Conf. Comput. Inf. Sci. ICCOINS 2016 - Proc.*, pp. 248–252, 2016, doi: 10.1109/ICCOINS.2016.7783222.

[15] K. Teknomo, *What is K-Nearest Neighbor Algoritm*. 2006.

[16] Kusrini and E. T. Lutfhi, *Algoritma Data Maining*. Yogyakarta, 2009.

[17] C. Vercellis, *Business Intelligent: Data Mining and Optimization for Decision Making*. Southern Gate, Chichester, West Sussex: John Willey & Sons, Ltd, 2009.

[18] R. Ndaumanu, "Analisis Prediksi Tingkat Pengunduran Diri Mahasiswa dengan Metode K-Nearest Neighbor," *Jatisi*, vol. 1, no. 3, 2004.