P-ISSN: 2655-3724 E-ISSN: 2720-9881
STATMAT (Jurnal Statistika dan Matematika)
Vol. 4, No. 2, Juli 2022
Artikel ke-7, Halaman: 166-177

# PERBANDINGAN UKURAN JARAK PADA ALGORITMA K-NEAREST NEIGHBOR DALAM ANALISIS SENTIMEN

**Alfiari Firdaus[1], Dwi Agustin Nuriani Sirodj[2*]**

[1]Program Studi Statistika, FMIPA, Universitas Islam Bandung
Email: falfiari@gmail.com
[2]Program Studi Statistika, FMIPA, Universitas Islam Bandung
*Email Korespondensi : dwi.agustinnuriani@unisba.ac.id

## *ABSTRACT*

*K-Nearest Neighbor (KNN) is one of the most widely used classification algorithms in machine learning methods. KNN classification is a conventional non-parametric classification method that has been used as a basic classifier in many pattern classification problems. The KNN search technique used in this study uses the Euclidean, Minkowski, Manhattan and linear least square distance formulas. The advantage of this method is that it is effective against noise data and is effective when the training data is large. However, this method still has drawbacks, namely the problem of the level of accuracy of the method used to measure the similarity between the objects being compared. The purpose of this study is to determine the application of the KNN method to sentiment analysis. The data used is 12,951 tweets taken from Twitter using the #OmicronVariant and #Covid19 hashtags. The results showed that the best k value parameter was 15. Using the Euclidean distance, the accuracy was quite good, and the recall was quite good, the precision was good, then the prediction results obtained that the positive category value was higher than the neutral category value and the negative category value. It can be concluded that the public's perception of Covid-19 Omicron is positive, meaning they believe in Omicron*
***Keywords: KNN, euclidean, minkowski, manhattan, linear least square.***

## ABSTRAK

*K-Nearest Neighbor* (KNN) merupakan salah satu algoritma klasifikasi yang paling banyak digunakan dalam metode *machine learning*. Klasifikasi KNN merupakan metode klasifikasi non-parametrik konvensional yang telah digunakan sebagai pengklasifikasi dasar dalam banyak masalah klasifikasi pola. Teknik pencarian KNN yang digunakan dalam penelitian ini dengan menggunakan rumus jarak *euclidean*, *minkowski*, *manhattan* dan *linear least square*. Keuntungan dari metode ini adalah efektif terhadap data *noise* dan efektif ketika data *training* berukuran besar. Namun, metode ini masih memiliki kekurangan yaitu masalah tingkat akurasi metode yang digunakan untuk mengukur kemiripan antar objek yang dibandingkan. Tujuan dari penelitian ini adalah untuk mengetahui ukuran jarak terbaik dalam metode KNN pada analisis sentimen. Data yang digunakan adalah data *tweet* sebanyak 12.951 yang diambil dari twitter dengan menggunakan *hastag #OmicronVariant* dan *#Covid19*. Hasil penelitian menunjukkan bahwa parameter nilai *k* terbaik adalah 15 sedangkan jarak terbaik adalah jarak *euclidean* yang diukur melalui nilai akurasi, *recall*, dan presisi yang baik, kemudian hasil prediksi diperoleh nilai kategori positif lebih tinggi dibandingkan nilai kategori netral dan nilai kategori negatif. Dapat disimpulkan bahwa persepsi masyarakat terhadap *Covid-19 Omicron* adalah positif, artinya mereka percaya dengan adanya virus covid-19 jenis o*micron.*

P-ISSN: 2655-3724 E-ISSN: 2720-9881
STATMAT (Jurnal Statistika dan Matematika)
Vol. 4, No. 2, Juli 2022
Artikel ke-7, Halaman: 166-177

**Kata kunci: KNN, *euclidean, minkowski, manhattan, linear least square.***

## 1. PENDAHULUAN

Di tengah kemajuan pesat inovasi penalaran terkomputerisasi (*Artificial Intelligence*) saat ini, kesadaran buatan manusia terdiri dari beberapa cabang, salah satunya adalah *machine learning*. *Machine learning* bertujuan untuk mengompilasi data yang diamati dari pengalaman yang dipelajari oleh program untuk menghasilkan informasi yang dapat dimanfaatkan (Mohamed, 2017). *K-Nearest Neighbor* (KNN) merupakan salah satu algoritma klasifikasi dalam metode *machine learning* yang paling banyak digunakan karena sederhana dan mudah diimplementasikan. Selain itu, biasanya digunakan sebagai pengklasifikasi dasar dalam banyak masalah domain (Jain et al., 2000).

Klasifikasi KNN adalah klasifikasi metode non-parametrik konvensional yang telah digunakan sebagai pengklasifikasi dasar dalam banyak masalah klasifikasi pola. Hal ini didasarkan pada pengukuran antara jumlah data testing dan data *training* untuk memutuskan klasifikasi akhir. Kelebihan metode KNN efektif terhadap data yang *noise* dan efektif apabila data *training* besar. Data *noise* yaitu *random error* atau varians dalam variable yang diukur, artinya terdapat kesalahan pada data yang bisa disebabkan oleh *human error* atau *outlier* yang menyimpang dari normal (Alasadi & Bhaya, 2017). Pada umumnya, metode pencarian KNN diselesaikan dengan menggunakan jarak *euclidean*. Jarak *euclidean* adalah formula untuk melacak jarak antara dua fokus dalam ruang dimensi dua. Dalam literatur, ada beberapa jenis fungsi jarak lainnya, seperti, *minkowski distance* (Batchelor, 1978), *manhattan distance*, dan *linear least square distance* (Singh et al., 2016). Dalam proses analisisnya, penggunaan teknik KNN untuk memutuskan jumlah *k* yang digunakan untuk mengkarakterisasi informasi baru. Besaran *k*, idealnya bilangan ganjil, misalnya $k = 1, 3, 5$, dan seterusnya. Kepastian nilai *k* dilihat berdasarkan seberapa banyak informasi yang ada dan ukuran aspek yang dibentuk oleh informasi tersebut. Semakin banyak informasi yang ada, semakin kecil jumlah *k* yang seharusnya diambil. Namun, semakin besar ukuran aspek informasi, semakin tinggi jumlah *k* yang harus diambil (Manning et al., 2009).

Dengan munculnya *web* 3.0 berbagai *platform* seperti *facebook*, *twitter*, *linkedIn*, *instagram* memungkinkan masyarakat untuk berbagi komentar, pandangan, perasaan, penilaian mereka tentang berbagai topik mulai dari pendidikan hingga hiburan. *Platform* ini berisi sejumlah besar data dalam bentuk *tweet*, *blog*, dan pembaruan *status*, *posting*, dan lain – lain. Banyaknya opini atau persepsi masyarakat di *platform* tersebut memunculkan berbagai tanggapan positif, negatif, atau bahkan netral. *Twitter* menjadi *platform* yang sering digunakan untuk mengungkapkan opini atau persepsi tentang berbagai hal. Dalam sehari *twitter* mampu menghasilkan jumlah *tweet* kurang lebih sebanyak 500 juta cuitan yang dikirimkan oleh penggunanya dari seluruh penjuru dunia (Syahnur et al., 2016). Sentimen adalah istilah yang digunakan untuk menggambarkan topik yang subjektif dan objektif dan topik faktual atau non-faktual yang melampaui perbedaan antara topik positif atau negatif (Pozzi et al., 2016). Analisis sentimen adalah pendekatan analitis yang digunakan untuk menganalisis sebuah teks. Tujuan dari analisis sentimen adalah untuk mengetahui subjektivitas opini, hasil *review* atau *tweet*. Berdasarkan analisis sentimen, opini dari seseorang dapat diklasifikasikan ke dalam berbagai kategori berdasarkan ukuran data dan jenis dokumen (Rajput et al., 2018). KNN merupakan metode pengklasifikasian, sehingga analisis sentimen dengan menggunakan metode KNN dapat menjadi solusi untuk menentukan hasil klasifikasi dari cuitan pada

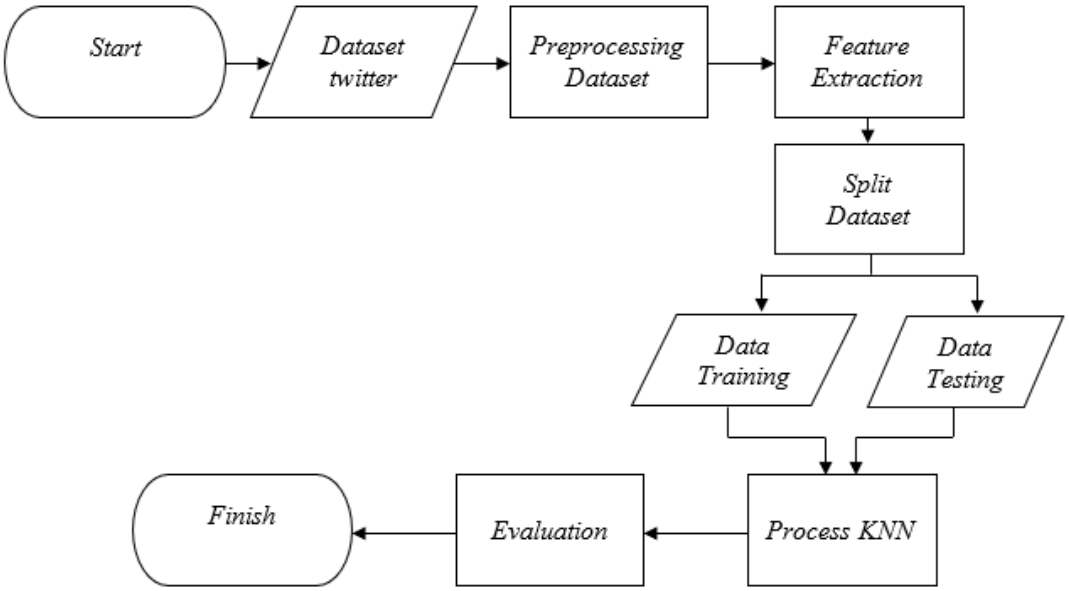
P-ISSN: 2655-3724 E-ISSN: 2720-9881
STATMAT (Jurnal Statistika dan Matematika)
Vol. 4, No. 2, Juli 2022
Artikel ke-7, Halaman: 166-177

*platform twitter*.

## 2. METODOLOGI

Penelitian ini menggunakan metode KNN. Populasi yang dipilih adalah cuitan *tweets* masyarakat Indonesia pada bulan Desember 2021 hingga Februari 2022 yang berjumlah 12.951 *tweets* yang diperoleh dari *website netlytics*. Variabel penelitian yang digunakan dalam penelitian ini adalah *tweet* atau opini yang dituangkan masyarakat dalam media sosial *twitter* dengan hashtag *#Covid19* dan *#OmicronVariant*.

Dalam pengambilan data *tweet* ini menggunakan teknik *crawling* sehingga diperoleh sebanyak 12.951 *tweets*. Pada penelitian ini akan diklasifikasikan data *tweets* tersebut dengan melihat tingkat akurasi, presisi, dan *recall* menggunakan *euclidean*, *minkowski*, *manhattan* dan *linear least square* dengan bantuan *software python*. Berikut merupakan *flowchart* untuk penelitian ini:




Gambar 1. *Flowchart* Analisis Penelitian

Dari Gambar 1 dapat dijelaskan beberapa tahapan diantaranya:

a. Input data *tweet*.
b. Melakukan *pre-processing* pada data *tweet*.
    (i) *Cleaning* data dengan cara menghapus terlebih dahulu akun-akun *bot* yang dilihat dari sisi *followers*, jam *post tweet*, *source tweet* yang tidak dikenal. Diasumsikan *followers* kurang dari 100 merupakan akun *bot*, jam *post tweet* dari jam 00.00 – 04.00 merupakan *bot*.
    (ii) Melakukan *case folding* yaitu mengubah kalimat yang didapat menjadi format yang sama dalam artian menjadi *lower case* semua.
    (iii) Melakukan tokenisasi yaitu menghilangkan *whitespace* dan membuang karakter tertentu seperti tanda baca, emoji dan url.
    (iv) Melakukan *stemming* yaitu menyederhanakan kata yang berisi imbuhan.
    (v) Melakukan normalisasi kata yaitu untuk mengurangi huruf berturut-turut dari suatu kata.

P-ISSN: 2655-3724 E-ISSN: 2720-9881
STATMAT (Jurnal Statistika dan Matematika)
Vol. 4, No. 2, Juli 2022
Artikel ke-7, Halaman: 166-177

(vi) Melakukan *stopword removal* yaitu menghilangkan kata umum yang sering muncul tetapi tidak memiliki arti penting dan tidak digunakan, contoh *has*, *and*, *he*, *being* dan sebagainya

c. Mengubah kalimat data *tweet* menjadi kategori, nilai 1 sebagai label sentimen positif, nilai 2 sebagai label sentimen negatif, dan nilai 3 sebagai label sentimen netral.

d. Membagi data *tweet* menjadi data *training* dan data *testing* dengan proporsi 80:20. Berdasarkan hasil penelitian oleh (Prakasa & Lhaksmana, 2018) dalam mengklasifikasikan data *tweet* menggunakan proporsi data training dan data testing sebesar 80:20 memberikan hasil akurasi yang paling baik yaitu 90.50% menggunakan metode KNN.

e. Menghitung jarak *euclidean*, *minkowski*, *manhattan* dan *linear least square* menggunakan *record* data *testing* dan data *training* menggunakan persamaan (1), (2), (3) dan (4).

$$d(i,j)_{euclidian} = \sqrt{\sum_{k=1}^{n}\left(x_{ik} - y_{jk}\right)^2} \quad (1)$$

Keterangan:
$d_{ij}$ = Jarak perhitungan kemiripan
$x_{ik}$ = Data *training*
$y_{jk}$ = Data *testing*

$$d(x,y)_{minkowski} = \left(\sum_{i=1}^{n}|x_i - y_i|^r\right)^{\frac{1}{r}} \quad (2)$$

Keterangan:
$d(x,y)$ = Jarak antara $x$ dan $y$
$n$ = Dimensi data
$x_i$ = Data *training* ke-$i$
$y_i$ = Data *testing* ke-$i$
$r$ = Parameter = 2

$$d(x,y)_{manhattan} = \sum_{i=1}^{m}|x_i - y_i| \quad (3)$$

Keterangan:
$x_i$ = Data *training* ke-$i$
$y_i$ = Data *testing* ke-$i$

$$d(x,y)_{linear\ least\ square} = \|y - xw\|^2 + \alpha * \|w\|^2 \quad (4)$$

Keterangan:
$x$ = data *training*
$y$ = data *testing*
$\alpha$ = parameter yang berfungsi mencegah *overfitting*
$w$ = parameter *linear least square* yang bisa digunakan tiga nilai yaitu (0.95, 0.001, 0.0004)

f. Setelah mendapatkan jarak *euclidean* selanjutnya menentukan jumlah *k* atau tetangga terdekat, *k* yang digunakan yaitu 1 sampai 40 dengan menggunakan *trial* and *error*. Semakin banyak dimensi yang ada, semakin rendah jumlah *k* yang seharusnya diambil. Namun, semakin besar ukuran aspek dimensi, semakin tinggi jumlah *k* yang harus diambil (Manning et al., 2009).

P-ISSN: 2655-3724 E-ISSN: 2720-9881
STATMAT (Jurnal Statistika dan Matematika)
Vol. 4, No. 2, Juli 2022
Artikel ke-7, Halaman: 166-177

g. Membuat *confusion matrix* untuk mengevaluasi penggunaan jarak dengan melihat tingkat akurasi, presisi dan recall menggunakan persamaan (5), (6), dan (7).

**Tabel 1.** *Confusion Matrix* yang terbentuk

| | | **Nilai Aktual** | | |
|---|---|---|---|---|
| | | ***Positive*** | ***Neutral*** | ***Negative*** |
| **Nilai Prediksi** | ***Positive*** | *True positive* (TP) | *False neutral* ($FNa_1$) | *False negative* ($FNe_1$) |
| | ***Neutral*** | *False positive* ($FP_1$) | *True neutral* (TNa) | *False negative* ($FNe_2$) |
| | ***Negative*** | *False positive* ($FP_2$) | *False neutral* ($FNa_2$) | *True negative* (TNe) |

Keterangan:
*TP* = Jumlah data positif yang terklasifikasi dengan benar oleh sistem
*TNe* = Jumlah data negatif yang terklasifikasi dengan benar oleh sistem
*FP* = Jumlah data positif yang terklasifikasi salah oleh sistem
*FNe* = Jumlah data negatif yang terklasfikasi salah oleh sistem
*TNa* = Jumlah data netral yang terklasifikasi dengan benar oleh sistem
*FNa* = Jumlah data netral yang terklasfikasi salah oleh sistem

$$\text{Accuracy} = \frac{\text{TP+TNa+TNe}}{\text{TP+TNa+TNe+FP}_1\text{+FP}_2\text{+FNa}_1\text{+FNa}_2\text{+FNe}_1\text{+FNe}_2} \times 100\% \quad (5)$$

$$\text{Precision}_+ = \frac{\text{TP}}{\text{TP+FP}_1\text{+FP}_2} \times 100\% \quad (6)$$

$$\text{Recall}_+ = \frac{\text{TP}}{\text{TP+FNa}_1\text{+FNe}_1} \times 100\% \quad (7)$$

Adapun pengkategorian tingkat akurasi adalah sebagai berikut (Gorunescu, 2011):
a) Nilai akurasi pada rentang 0,90 – 1,00 = *excellent classification*
b) Nilai akurasi pada rentang 0,80 – 0,90 = *good classification*
c) Nilai akurasi pada rentang 0,70 – 0,80 = *fair classification*
d) Nilai akurasi pada rentang 0.60 – 0,70 = *poor classification*
e) Nilai akurasi pada rentang 0.50 – 0.60 = *failure*

## 3. HASIL DAN PEMBAHASAN

### 3.1. Preprocessing Data

*Preprocessing* data dilakukan dengan beberapa tahapan yaitu *case folding*, tokenisasi, *stemming*, normalisasi kata, dan *stopwords removal* sehingga menghasilkan data yang bersih dan bisa dilanjutkan untuk proses selanjutnya.

a. *Case Folding*

P-ISSN: 2655-3724 E-ISSN: 2720-9881
STATMAT (Jurnal Statistika dan Matematika)
Vol. 4, No. 2, Juli 2022
Artikel ke-7, Halaman: 166-177

*Case folding* disini adalah proses mengubah data *tweet* menjadi *lowercase*. Pada Gambar 2 berikut merupakan contoh proses *case folding*:

| *Tweet* | *Case Folding* |
|---|---|
| "The #OmicronVariant , #COVID19, mandatory #masks and #vaccination; A thread of hope. It has been a hard two years. … | "the #omicronvariant , #covid19, mandatory #masks and #vaccination; a thread of hope. it has been a hard two years. … |
| "☐ Operations for International arrivals are running smooth after the implementation of the new guidelines laid down by Ministry of Health … | "☐ operations for international arrivals are running smooth after the implementation of the new guidelines laid down by ministry of health … |
| … | … |
| … | … |

Gambar 2. Proses *Case Folding*

Dapat dilihat pada Gambar 2. bahwa seluruh kalimat *tweet* yang mulanya terdapat huruf *uppercase* menjadi *lowercase*. Sebagai contoh pada kalimat *tweet* awal "The #OmicronVariant…" masih mengandung huruf kapital setelah dilakukan proses *case folding* seluruh kalimat *tweet* sudah tidak mengandung huruf kapital.

b. Tokenisasi

Tokenisasi disini merupakan proses memecah *string* atau *input* terhadap suatu teks yang telah melewati proses *case folding* berdasarkan tiap kata dan menghilangkan *url*, *@mention*, dan *hashtag*. Pada Gambar 3 berikut merupakan contoh proses tokenisasi:

| *Tweet* | **Tokenisasi** |
|---|---|
| "🚨 🚨 discussing #covid19 with @iromg ▪ masks back in england but cases in wales higher with masks already in force… | discussing with masks back in england but cases in wales higher with masks already in force… |
| "sigh of relief in south africa as omicron variant appears to be 'a super mild' mutation who … | sigh of relief in south africa as omicron variant appears to be a super mild mutation who … |
| "while the new covid-19 variant has everyone spooked, cautious minds are waiting for … | while the new covid19 variant has everyone spooked cautious minds are waiting for … |
| "the #omicronvariant , #covid19, mandatory #masks and #vaccination; a thread of hope. it has been a hard two years. … | The mandatory and a thread of hope it has been a hard two years … |
| "☐ operations for international arrivals are running smooth after the implementation of the new guidelines laid down by ministry of health … | operations for international arrivals are running smooth after the implementation of the new guidelines laid down by ministry of health … |
| … | … |

Gambar 3. Proses Tokenisasi

Pada Gambar 3 terdapat perbedaan antara kolom *tweet* dengan kolom tokenisasi. Pada kolom tokenisasi kalimat *tweet* tidak mengandung unsur *hastag*, *@mention*, dan *emoticon*.

c. *Stemming*

P-ISSN: 2655-3724 E-ISSN: 2720-9881
STATMAT (Jurnal Statistika dan Matematika)
Vol. 4, No. 2, Juli 2022
Artikel ke-7, Halaman: 166-177

Proses ini adalah mencari kata dasar dari setiap kata dari hasil proses preprocessing sebelumnya dengan menghilangkan kata imbuhan baik didepan maupun dibelakang kata. Pada Gambar 4 berikut merupakan contoh proses *stemming*:

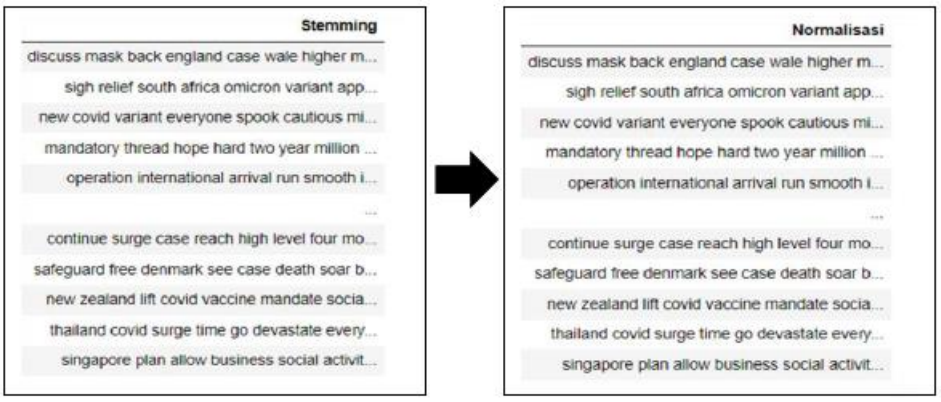
| *Tweet* | *Stemming* |
|---|---|
| discussing with masks back in england but cases in wales higher with masks already in force… | discuss with mask back in England but case wales high mask already in force … |
| sigh of relief in south africa as omicron variant appears to be a super mild mutation who … | sigh of relief in south africa as omicron variant appear to be a super mild mutation who … |
| while the new covid19 variant has everyone spooked cautious minds are waiting for … | while the new covid19 variant has everyone spooke cautious minds are wait for … |
| The mandatory and a thread of hope it has been a hard two years … | The mandatory and a thread of hope it has been a hard two year … |
| operations for international arrivals are running smooth after the implementation of the new guidelines laid down by ministry of health … | operation for international arrival are run smooth after the implementation of the new guideline laid down by ministry of health … |
| … | … |

Gambar 4. Proses *Stemming*

Pada Gambar 4 dapat dilihat bahwa pada kolom *tweet* data pertama semula terdapat kata *discussing* menjadi *discuss* begitupun dengan kata lainnya. Hal ini membuktikan bahwa proses *stemming* akan mencari kata dasar dari seluruh kalimat data *tweets*. Selain untuk mengubah menjadi kata dasar, juga untuk melakukan pengelompokan kata-kata lain yang memiliki kata dasar dan arti yang serupa namun memiliki bentuk yang berbeda karena mendapatkan imbuhan yang berbeda pula.

d. Normalisasi Kata

Proses ini merupakan proses untuk mengurangi huruf yang berlebih dalam suatu kata. Pada Gambar 5 contoh proses normalisasi kata:




Gambar 5. Proses Normalisasi Kata

Pada Gambar 5 dijelaskan proses normalisasi kata yaitu terdapat kata *environmenttt* menjadi *environment*. Tujuan proses ini adalah untuk menghilangkan redundansi data (pengulangan) dan menstandarisasi informasi untuk alur kerja data yang lebih baik.

e. *Stopwords Removal*

P-ISSN: 2655-3724 E-ISSN: 2720-9881
STATMAT (Jurnal Statistika dan Matematika)
Vol. 4, No. 2, Juli 2022
Artikel ke-7, Halaman: 166-177

Proses ini merupakan proses menghilangkan kata-kata yang dianggap tidak perlu untuk proses selanjutnya seperti pada Gambar 6.

```
print(stopwords.words('english'))
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're",
"you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he',
'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "i
t's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what',
'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'i
s', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'havin
g', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or',
'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'a
gainst', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'b
elow', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under',
'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'h
ow', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'suc
h', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very',
's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now',
'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "could
n't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't",
'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "must
n't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wa
sn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
```

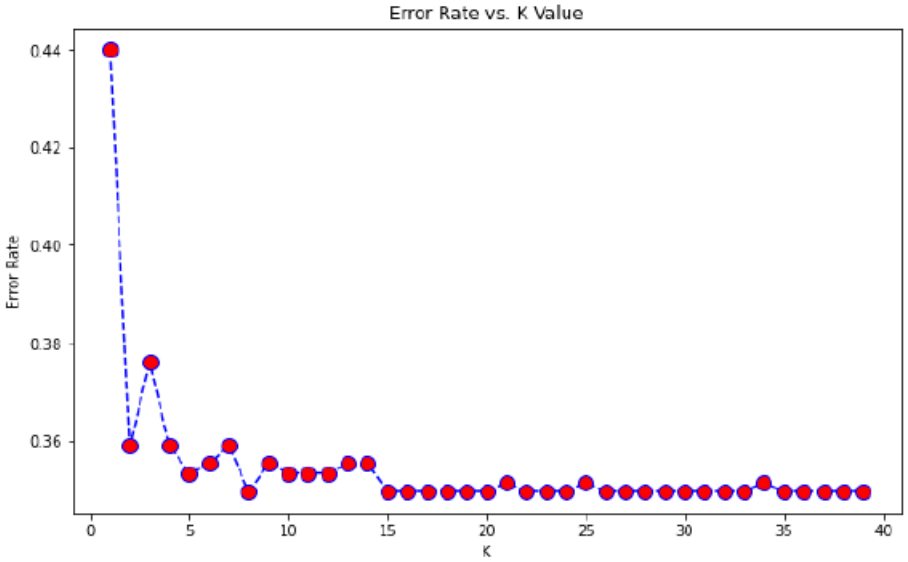
Gambar 6. *Stopwords* dalam Bahasa Inggris

Dalam Gambar 6, dapat dilihat bahwa terdapat *stopwords* dalam Bahasa Inggris. Apabila dalam kalimat *tweet* terdapat *stopwords* diatas maka akan dihilangkan untuk proses atau tahapan selanjutnya. Namun *stopwords* diatas masih bisa ditambahkan jika kebutuhan peneliti masih dirasa kurang.

### 3.2. Implementasi KNN pada Analisis Sentimen

a. *Euclidean Distance*

Proses *euclidean distance* dimulai dengan menentukan nilai *k* dengan menggunakan data *training* kemudian dilihat dari nilai *error rate*. Berikut merupakan grafik dari nilai *error rate*nya:

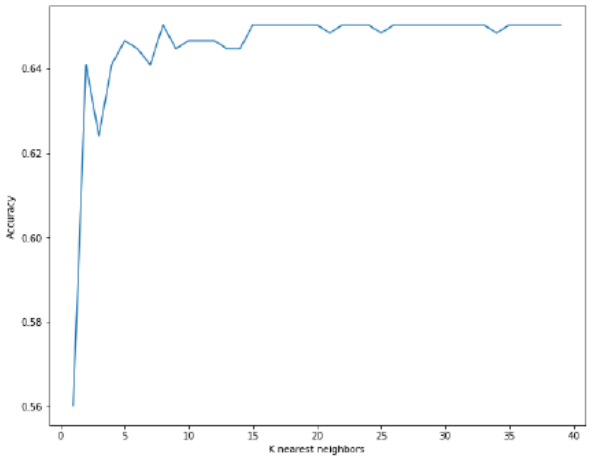



Gambar 7. Proses Penentuan Nilai *k* dari nilai *error rate*

Pada Gambar 7, pada nilai $k \geq 15$ nilai *error rate* cenderung konvergen yang

P-ISSN: 2655-3724 E-ISSN: 2720-9881
STATMAT (Jurnal Statistika dan Matematika)
Vol. 4, No. 2, Juli 2022
Artikel ke-7, Halaman: 166-177

diperoleh nilai *error rate* kurang dari 0.36. Maka dapat ditarik kesimpulan untuk penentuan nilai *k* yang diambil yaitu *k* = 15. Kemudian akan dibuktikan kembali menggunakan data *training* dengan melihat nilai akurasi dari penggunaan nilai *k* yang diperoleh yaitu 15. Berikut merupakan grafik nilai akurasinya:




Gambar 8. Nilai akurasi dengan memasukan nilai *k*

Pada Gambar 8 diperoleh bahwa, apabila nilai $k \geq 15$ maka akan diperoleh nilai akurasi lebih dari 0.64 dan cenderung konvergen. Maka dapat disimpulkan kembali bahwa penggunaan nilai *k* = 15 akan memperoleh nilai akurasi yang cukup tinggi atau dengan menggunakan data *training* dan nilai *k* = 15, nilai akurasi yang didapat yaitu 64.12%. Proses berikutnya yaitu menggunakan data *testing*. Dengan menggunakan nilai *k* = 15 diperoleh nilai akurasi sebesar 65.04%. Maka setelah diketahui nilai *k* dan nilai akurasi menggunakan *euclidean distance* tahapan selanjutnya akan menggunakan ketiga fungsi jarak lainnya yaitu *minkowski distance*, *manhattan distance*, dan *linear least square distance*.

Penerapan fungsi jarak pada analisis sentimen kasus *covid-19 omicron* yaitu untuk melihat pendapat atau kecenderungan opini seseorang, apakah cenderung beropini positif, negatif, atau netral. Serta dapat memberikan gambaran akurasi yang didapat dari analisis sentimen tersebut. Berikut merupakan hasil *confusion matrix* menggunakan jarak *euclidean distance* dengan nilai *k* = 15:

Tabel 2. *Confusion Matrix Euclidean Distance* dengan *k* = 15

| | | **Nilai Aktual** | | |
|---|---|---|---|---|
| | | *Positive* | *Neutral* | *Negative* |
| **Nilai** | *Positive* | 344 | 0 | 2 |
| **Prediksi** | *Neutral* | 91 | 0 | 1 |
| | *Negative* | 92 | 0 | 2 |

Pada Tabel 2 diperoleh prediksi dari penggunaan jarak *euclidean* yaitu prediksi terhadap cuitan sentiment kategori positif yang diklasifikasikan dengan benar sebagai sentiment kategori positif ada sebanyak 344 tetapi prediksi terhadap cuitan sentiment kategori netral yang diklasifikasikan dengan benar sebagai

P-ISSN: 2655-3724 E-ISSN: 2720-9881
STATMAT (Jurnal Statistika dan Matematika)
Vol. 4, No. 2, Juli 2022
Artikel ke-7, Halaman: 166-177

sentimen kategori netral ada sebanyak 0 dan prediksi terhadap cuitan sentiment kategori negatif yang diklasifikasikan dengan benar sebagai sentiment kategori negatif ada sebanyak 2. Selanjutnya dari hasil *confusion matrix* dapat dihitung nilai akurasi, presisi dan *recall*. Berikut merupakan hasilnya:

$$\text{Accuracy} = \frac{344 + 0 + 2}{344 + 0 + 2 + 91 + 92 + 0 + 0 + 2 + 1} \times 100\% = 65.04\%$$

$$\text{Precision}_{+} = \frac{344}{344 + 91 + 92} \times 100\% = 65.28\%$$

$$\text{Recall}_{+} = \frac{344}{344 + 0 + 2} \times 100\% = 99.42\%$$

Dengan menggunakan jarak *euclidean* pada data *tweets* diperoleh nilai akurasi sebesar 65.04% menandakan model kurang baik dalam melakukan klasifikasi (*poor classification*). Nilai presisi sebesar 65.28% menandakan persentase cuitan *tweets* yang benar masuk kategori sentiment positif dari keseluruhan cuitan *tweets* yang diprediksi sentiment positif dan nilai *recall* sebesar 99.42% menandakan persentase cuitan *tweets* positif yang diprediksi sentiment positif dibandingkan keseluruhan cuitan *tweets* yang sebenarnya masuk kedalam kategori sentiment positif.

b. Komparasi Keempat Jarak

Setelah diperoleh nilai akurasi menggunakan nilai $k$ = 15, selanjutnya akan dibandingkan untuk melihat ukuran jarak mana yang terbaik untuk data *tweets* pada analisis sentiment. Berikut merupakan tabel hasil perhitungan akurasi, presisi dan *recall* pada masing – masing jarak:

Tabel 3. Komparasi Keempat Jarak dengan melihat nilai Akurasi

| Nama Jarak | Nilai *k* | Nilai Akurasi | Nilai Presisi | Nilai *Recall* |
|---|---|---|---|---|
| *Euclidean* | 15 | **65.04%** | 65.28% | **99.42%** |
| Minkowski | 15 | 64.66% | **65.33%** | 98.85% |
| Manhattan | 15 | 64.85% | 65.03% | **99.42%** |
| *Linear Least Square* | 15 | 64.54% | 65.09% | 98.85% |

Pada Tabel 3, dalam menentukan jarak terbaik dapat dilihat berdasarkan nilai akurasi, presisi, dan *recall* tertinggi. Dari keempat jarak tersebut dapat dilihat bahwa jarak *euclidean* memiliki nilai akurasi tertinggi yaitu sebesar 65.04% dan nilai *recall* tertinggi sebesar 99.42%. Sedangkan berdasarkan nilai presisi jarak *minkowski* memiliki nilai presisi tertinggi yaitu sebesar 65.33%. Sehingga dapat disimpulkan ukuran jarak yang cocok untuk jenis data *tweets* yaitu jarak *euclidean*. Adapun hasil perhitungan menggunakan jarak *euclidean* prediksi sentiment yang masuk kedalam kategori positif, negatif, dan netral sebagai berikut:

P-ISSN: 2655-3724 E-ISSN: 2720-9881
STATMAT (Jurnal Statistika dan Matematika)
Vol. 4, No. 2, Juli 2022
Artikel ke-7, Halaman: 166-177

Tabel 4. Persentase Nilai Prediksi *Euclidean Distance* dengan $k$ = 15

| Kategori Prediksi | Nilai Prediksi | Persentase |
|---|---|---|
| **Positif** | 346 | 65.04% |
| **Negatif** | 92 | 17.67% |
| **Netral** | 94 | 17.29% |
| **Total** | 532 | 100% |

Pada Tabel 4 menjelaskan nilai persentase prediksi jarak *euclidean* menggunakan nilai $k$ = 15 terhadap analisis sentiment data *tweets* dengan *hashtag #OmicronVariant* dan *#Covid19* yaitu diperoleh persentase kategori sentiment positif sebesar 65.04% artinya *tweets* tersebut mempercayai adanya *omicron*, sedangkan kategori sentiment negatif sebesar 17.67% artinya *tweets* tersebut kontra atau tidak mempercayai adanya *omicron* dan persentase kategori sentiment netral sebesar 17.29% artinya *tweets* tersebut hanya sebuah berita atau tidak pro maupun kontra terhadap *omicron*.

## 4. SIMPULAN

Penerapan metode KNN pada analisis sentiment berhasil dilakukan dengan menggunakan keempat jarak yaitu *euclidean distance*, *minkowski distance*, *manhattan distance*, dan *linear least square distance*. Dengan menggunakan nilai $k$ = 15, diperoleh hasil akurasi tertinggi menggunakan jarak *euclidean* dengan nilai akurasi sebesar 65.04% yang berarti model cukup baik dalam melakukan klasifikasi kemudian nilai presisi sebesar 65.28% menandakan persentase cuitan *tweets* yang benar masuk kategori sentiment positif dari keseluruhan cuitan tweets yang diprediksi sentiment positif dan nilai *recall* sebesar 99.42% artinya persentase cuitan *tweets* positif yang diprediksi sentiment positif dibandingan keseluruhan cuitan *tweets* yang sebenarnya masuk kedalam kategori sentiment positif. Maka penerapan metode KNN pada analisis sentiment menggunakan data *tweets* akan cocok jika menggunakan jarak *euclidean*.

## 5. DAFTAR PUSTAKA

Alasadi, S. A., & Bhaya, W. S. (2017). Review of Data Preprocessing Techniques in Data Mining. *Journal of Engineering and Applied Sciences*, *12*(16), 4102–4107. https://doi.org/10.3923/jeasci.2017.4102.4107

Batchelor, B. G. (1978). *Pattern Recognition Ideas in Practice* (1st ed. 19). New York, NY : Springer US : Imprint: Springer.

Gorunescu, F. (2011). *Data Mining: Concepts, Models and Techniques*. https://doi.org/10.1007/978-3-642-19721-5

Jain, A. K., Duin, R. P., & Mao, J. (2000). Statistical Pattern Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(1), 4–37. https://doi.org/10.1109/34.824819

Manning, C. D., Raghavan, P., & Schutze, H. (2009). An Introduction to Modern Information Retrieval. In *Cambridge University Press* (Online Edi, Vol. 53, Issue 9). Cambridge University Press. https://doi.org/10.1108/00242530410565256

Mohamed, A. E. (2017). Comparative Study of Machine Learning Techniques for Supervised Classification of Biomedical Data. *International Journal of Applied Science and Technology*, *7*(2), 5–18. https://doi.org/10.15546/aeei-2014-0021

Pozzi, F. A., Fersini, E., Messina, E., & Liu, B. (2016). *Sentiment Analysis in Social*

P-ISSN: 2655-3724 E-ISSN: 2720-9881
STATMAT (Jurnal Statistika dan Matematika)
Vol. 4, No. 2, Juli 2022
Artikel ke-7, Halaman: 166-177

*Networks* (F. A. Pozzi, E. Fersini, E. Messina, & B. Liu, Eds.; 1st editon, pp. 1–284). Morgan Kaufmann. https://doi.org/10.1016/C2015-0-01864-0

Prakasa, O. S. Y., & Lhaksmana, K. M. (2018). Klasifikasi Teks Dengan Menggunakan Algoritma K-nearest Neighbor Pada Kasus Kinerja Pemerintah Di Twitter. *EProceedings of Engineering*, *5*(3), 8237–8248.

Rajput, D. S., Thakur, R. S., & Basha, S. M. (2018). *Sentiment Analysis and Knowledge Discovery in Contemporary Business* (D. S. Rajput, R. S. Thakur, & S. M. Basha, Eds.; pp. 1–333). IGI Global. https://doi.org/10.4018/978-1-5225-4999-4

Singh, J., Singh, G., & Singh, R. (2016). A review of sentiment analysis techniques for opinionated web text. *CSI Transactions on ICT*, *4*(2–4), 241–247. https://doi.org/10.1007/s40012-016-0107-y

Syahnur, M. H., Bijaksana, M. A., & Mubarok, M. S. (2016). Kategorisasi Topik Tweet di Kota Jakarta, Bandung, dan Makassar dengan Metode Multinomial Naïve Bayes Classifier. *E-Proceeding of Engineering*, *3*(2), 3612–3620.