

4267-11164-1-SM

by Faiza R Irawan

Submission date: 21-Mar-2022 07:35PM (UTC+0700)

Submission ID: 1789190561

File name: 4267-11164-1-SM.docx (998.76K)

Word count: 3743

Character count: 23269

ANALISIS SENTIMEN TERHADAP PENGGUNA GOJEK MENGGUNAKAN METODE K-NEAREST NEIGHBORS

Faiza R Irawan¹, Ahmad Jazuli², Tutik Khotimah³

¹Program Studi Teknik Informatika

²Universitas Muria Kudus

³Universitas Muria Kudus

Email: ¹201851234@std.umk.ac.id, ²ahmad.jazuli@umk.ac.id, ³tutik.khotimah@umk.ac.id

(Naskah masuk: dd mmm yyyy, diterima untuk diterbitkan: dd mmm yyyy)

Abstrak

Zaman modern menjadikan internet sebagai hal wajar, masyarakat sering menggunakan media sosial sebagai *microblogging*. Masyarakat menggunakannya untuk mengekspresikan diri dan mengemukakan pendapat. Salah satu media sosial yang digunakan yaitu twitter. Informasi yang dapat dibagikan pada twitter dinamakan *tweet*. *Tweet* memiliki batas yang dulunya 140 karakter menjadi 280 karakter. *Tweet* dapat berupa teks, foto, video, dan sebagainya, melalui *tweet* pengguna dapat berinteraksi satu sama lain. Mereka dapat membagikan informasi, memberikan pendapat, menjelaskan suatu kejadian dan sebagainya. Gojek merupakan *startup* yang bergerak dibidang *ride hailing*. Perusahaan Gojek sudah berdiri dari tahun 2010 oleh Nadiem Makarim, hampir seluruh kota di Indonesia sudah terjangkau Gojek. Setiap tahunnya dibutuhkan evaluasi dan juga penilaian terhadap layanan Gojek untuk mengetahui respon pengguna Gojek. Penelitian ini membahas klasifikasi sentimen positif, negatif, atau netral terhadap respon pengguna Gojek di media sosial twitter dengan menerapkan metode *K-Nearest Neighbor*. Hasil klasifikasi menggunakan metode *K-Nearest Neighbor* mampu mengklasifikasikan respon dari pengguna twitter dan dapat digunakan perusahaan Gojek sebagai bahan evaluasi dan penilaian terhadap layanan Gojek. Hasil pengujian metode *K-Nearest Neighbor* menggunakan *confusion matrix* dengan data sebanyak 1409 mendapatkan tingkat akurasi sebesar 79,43% dengan nilai $k=15$.

Kata kunci: Analisis Sentimen, confusion matrix, Gojek, K-Nearest Neighbor, Twitter.

SENTIMENT ANALYSIS OF GOJEK USERS USING K-NEAREST NEIGHBOR METHOD

Abstract

Modern times make the internet a natural thing, people often use social media as *microblogging*. People use it to express themselves and express opinions. One of the social media used is Twitter. Information that can be shared on Twitter is called a *tweet*. *Tweets* have a limit from 140 characters to 280 characters. *Tweets* can be text, photos, videos, and so on, through *tweets* users can interact with each other. They can share information, give opinions, explain an incident and so on. Gojek is a *startup* engaged in *ride-hailing*. The Gojek company has been established since 2010 by Nadiem Makarim, almost all cities in Indonesia have been reached by Gojek. Every year an evaluation and assessment of Gojek's services is needed to find out the response of Gojek users. This study discusses the classification of positive, negative, or neutral sentiments on the response of Gojek users on Twitter social media by applying the *K-Nearest Neighbor* method. The results of the classification using the *K-Nearest Neighbor* method are able to classify responses from twitter users and can be used by the Gojek company as an evaluation and assessment material for Gojek services. The results of applying the *K-Nearest Neighbor* method using a *confusion matrix* with 1409 data get an accuracy rate of 79.43% with a value of $k=15$.

Keywords: Sentiment Analysis, confusion matrix, Gojek, K-Nearest Neighbor, Twitter.

1. PENDAHULUAN

Masyarakat sering menggunakan media sosial sebagai *microblogging*. Masyarakat menggunakannya sebagai ungkapan ekspresikan diri ataupun mengemukakan pendapat [1]. Salah satu media

sosial yang digunakan yaitu twitter. Informasi yang dapat dibagikan pada twitter dinamakan *tweet*. *Tweet* memiliki batas yang dulunya 140 karakter menjadi 280 karakter. *Tweet* dapat berupa teks, foto, video, dan sebagainya, melalui *tweet* pengguna dapat berinteraksi satu sama lain. Mereka dapat

membagikan informasi, memberikan pendapat, menjelaskan suatu kejadian dan sebagainya.

Gojek merupakan startup yang bergerak dibidang *ride hailing*, Perusahaan Gojek sudah berdiri dari tahun 2010 oleh Nadiem Makarim, hampir seluruh kota di Indonesia sudah terjangkau Gojek. Setelah perkembangan yang cukup pesat, Gojek melebarkan sayapnya ke sektor lain seperti GoRide, GoCar, GoBluebird, GoFood, GoMart, GoMall, GoSend dan masih banyak lagi, hal ini menjadikan gojek sebagai salah satu startup terbesar di Indonesia. Aplikasi gojek telah diunduh sebanyak 190 juta lebih, dengan 2 juta lebih *driver* yang telah terdaftar, dan 900 rb lebih pengguna yang telah menjadi Mitra.

Perusahaan gojek melakukan pengembangan sistem berdasarkan keluhan atau masukan dari pengguna untuk memperbaiki layanan gojek agar lebih baik. Twitter merupakan salah satu platform sosial media untuk mendapatkan informasi yang dapat digunakan sebagai pedoman evaluasi dan penilaian su⁴³ layanan, dalam hal ini dibutuhkan pendekatan analisis sentimen dengan menggunakan metode *K-Nearest Neighbor* dalam mengklasifikasi senti⁴⁴n positif, negatif, dan netral.

K-Nearest Neighbor (KNN) termasuk dalam algoritma *supervised learning*. Algoritma ini banyak diaplikasikan pada data mining, *statistical pattern recognition*, *image processing*. Algoritma *K-Nearest Neighbor* (KNN) dalam memetakan suatu objek berdasarkan atribut dan sampel dari data training sehingga sangat efektif apabila menggunakan data training yang besar untuk hasil lebih akurat. Hal ini diharapkan mampu memberikan informasi tentang respon pengguna gojek terhadap layanan yang sudah diberikan oleh gojek yang nantinya dapat digunakan sebagai pedoman dalam penilaian dan evaluasi layanan gojek.

2. METODE PENELITIAN

a. Analisis Sentimen

Tujuan dari analisis sentimen adalah untuk menilai emosi, persepsi, dan penilaian konsumen atau peneliti terhadap produk, layanan, atau kepribadian seseorang[1]. Beberapa studi tentang produk didahului dengan menentukan pembahasan suatu produk sebelum memulai proses eksplorasi opini. *Sentiment analysis* adalah proses memahami dan mengelompokkan apa yang ditemukan dalam sebuah tulisan melalui penggunaan teknik analisis teks. Analisis sentimen sering disebut sebagai "*opinion mining*"[2]. Ini menunjukkan bahwa itu akan memunculkan emosi yang tersembunyi dibalik kata-kata setiap pelanggan. Saat ini, pelanggan senang menggunakan platform seperti media sosial, *e-commerce*, dan situs web untuk mengekspresikan diri. Salah satu keuntungan dari analisis sentimen adalah dapat dilakukan secara otomatis, sehingga menghemat waktu dan sumber daya. Berbagai alat

telah dikembangkan oleh algoritma khusus untuk analisis data skala besar. Detail dari analisis sentimen dalam n³¹klasifikasikan sentimen positif, negatif, dan netral dapat dilihat pada Tabel 1.

Tabel 1. Analisis Sentimen

No.	Sentimen	Keterangan
1.	Positif	Sentimen positif merupakan respon ataupun opini seseorang yang menandakan dirinya setuju akan suatu keadaan.
2.	Netral	Sentimen netral merupakan respon atau opini yang tidak bersifat negatif ataupun positif.
3.	Negatif	Sentimen negatif merupakan respon ataupun opini seseorang yang menandakan dirinya bertentangan, marah, tidak setuju dan sebagainya, biasanya sentimen negatif ditandai dengan negasi.

b. Text Mining

⁸ *Text mining* adalah suatu proses pengumpulan informasi secara intensif dengan menggunakan alat dan metode tertentu, ini dapat digunakan untuk menganalisis suatu data ataupun dokumen. *Text mining* termasuk dalam data mining, dimana *text mining* dapat menganalisis data semi terstruktur (Word, PDF, kutipan teks) dan tidak terstruktur. Tujuan *text mining* adalah untuk mer¹⁶ami dan memperoleh informasi yang bermanfaat dari sumber data dengan mengidentifikasi dan mengeksplorasi pola linguistik yang unik. Dalam kasus *text mining*, sumber data yang digunakan adalah *collection* atau *unstructured collection* dan memerlukan kategorisasi untuk menemukan informasi sejenis. *Text mining* diimplementasikan dalam *spamming*, analisis sentimen, pengukuran preferensi pelanggan, sintesis dokumen, dan sebagainya[3].

c. Gojek

Gojek merupakan *startup* yang bergerak dibidang *ride hailing*, Perusahaan Gojek sudah berdiri dari tahun 2010 oleh Nadiem Makarim, hampir seluruh kota di Indonesia sudah terjangkau Gojek[4]. Setelah perkembangan yang cukup pesat, Gojek melebarkan sayapnya ke sektor lain seperti GoRide, GoCar, GoBluebird, GoFood, GoMart, GoMall, GoSend dan masih banyak lagi, hal ini menjadikan Gojek sebagai salah satu startup terbesar di Indonesia. Aplikasi Gojek telah diunduh sebanyak 190 juta lebih, dengan 2 juta lebih *driver* yang telah terdaftar, dan 900 rb lebih pengguna yang telah menjadi Mitra.

d. Metode K-Nearest Neighbor

Metode *K-Nearest Neighbor* adalah salah satu metode *supervised learning* yang proses klasifikasinya berdasarkan jarak ke⁴⁵nggaan terdekat dari dataset training [5]. *K-Nearest Neighbor* termasuk metode yang sering digunakan untuk mengimplementasikan perhitungan jarak

dalam proses klasifikasi karena memiliki rumus yang sederhana [6]. Adapun kelebihan metode ini adalah memiliki ketahanan terhadap data training yang mempunyai banyak *noise* serta efektif digunakan untuk data training yang besar. Selain itu, metode *K-Nearest Neighbor* juga memiliki beberapa kelemahan seperti belum jelasnya jenis perhitungan jarak dan yang harus digunakan untuk memperoleh hasil terbaik, memerlukan biaya komputasi yang cukup tinggi, dan perlunya menentukan nilai dari parameter *k* [7]. Adapun rumus persamaan jarak *Euclidean* dapat dilihat pada persamaan dibawah ini.

$$d_i = \sqrt{\sum_{i=1}^n (X_{2i} - X_{1i})^2} \quad (1)$$

Keterangan :

- d_i = jarak *euclidean* ke-i
- X_{2i} = jarak *training* ke-i
- X_{1i} = jarak *testing* ke-i
- n = banyak data *training*
- I = baris ke -i

e. Twitter

Twitter menjadi bagian dari *Computer Mediated Communication* (komunikasi manusia melalui komputer yang melibatkan banyak orang), proses ini memanfaatkan media tertentu. Pengguna twitter mewakili berbagai macam pengguna sehingga memudahkan dalam berinteraksi dengan teman, keluarga, ataupun rekan kerja. Twitter memberikan akses kepada penggunaannya untuk mengirimkan pesan singkat (*tweet*) dengan maksimal 140 karakter menjadi 280 karakter [8]. Tweet dapat berupa teks, foto, video dan sebagainya, melalui *tweet* pengguna dapat berinteraksi satu sama lain. Mereka dapat membagikan informasi, memberikan pendapat, menjelaskan suatu kejadian dan sebagainya.

f. Alur Penelitian

Untuk menapai tujuan dalam penelitian ini ada beberapa tahapan-tahapan diantaranya:

1. Metode Pengambilan Data

Pengambilan data bertujuan untuk mengumpulkan *dataset* pada penelitian ini. Data yang akan digunakan berupa respon pengguna gojek di media sosial twitter melalui twitter API yaitu *tweet* berbahasa Indonesia dengan kata kunci gojek. Berikut merupakan diagram alir metode pengambilan data dapat dilihat pada gambar 1:



Gambar 1. Metode pengambilan data

2. Metode Pengolahan Data

Dataset yang sudah didapatkan melalui proses *crawling* dengan memanfaatkan twitter API

merupakan dataset mentah. Dataset tersebut perlu diolah sebelum diklasifikasi dengan metode *K-Nearest Neighbor*, proses ini disebut *preprocessing*. Pada *preprocessing* ada lima tahap yaitu:

1) *Cleaning*

Tahap *cleaning* bertujuan untuk menghapus simbol, tanda baca, angka, url, *mention* dan menghasilkan kalimat yang akan diproses ke tahap selanjutnya[9]. Tahap *cleaning* dapat dilihat pada tabel 2.:

Tabel 2. *Cleaning*

<i>Tweet</i>	<i>Cleaning</i>
Bahkan pas di Bandung pun beli oleh oleh via gojek kirim ke hotel karena promo gofood gede bangett ☐ bisa nyenengin semua orang temen kantor ayah temen mama tetangga dengan biaya hemat ☐ makasih gojek gofood ♥ #CepetanGoFood https://t.co/DENBeduDth	Bahkan pas di Bandung pun beli oleh oleh via gojek kirim ke hotel karena promo gofood gede bangett bisa nyenengin semua orang temen kantor ayah temen mama tetangga dengan biaya hemat makasih gojek gofood CepetanGoFood

2) *Case Folding*

Case folding merubah dokumen huruf besar menjadi lowecase atau huruf kecil[10]. Tahap *case folding* dapat dilihat pada tabel 3.:

Tabel 3. *Case Folding*

<i>Tweet</i>	<i>Case Folding</i>
Bahkan pas di Bandung pun beli oleh oleh via gojek kirim ke hotel karena promo gofood gede bangett ☐ bisa nyenengin semua orang temen kantor ayah temen mama tetangga dengan biaya hemat ☐ makasih gojek gofood ♥ #CepetanGoFood https://t.co/DENBeduDth	bahkan pas di bandung pun beli oleh oleh via gojek kirim ke hotel karena promo gofood gede bangett bisa nyenengin semua orang temen kantor ayah temen mama tetangga dengan biaya hemat makasih gojek gofood cepetangofood

3) *Tokenizing*

Tokenizing digunakan untuk memisahkan jarak kata dan spasi untuk digunakan pembobotan data[11]. Tahap *Tokenizing* dapat dilihat pada tabel 4.:

Tabel 4. *Tokenizing*

<i>Tweet</i>	<i>Tokenizing</i>
Bahkan pas di Bandung pun beli oleh oleh via gojek kirim ke hotel karena promo gofood gede bangett ☐ bisa nyenengin semua orang temen kantor ayah temen mama tetangga dengan biaya hemat ☐ makasih gojek gofood ♥ #CepetanGoFood https://t.co/DENBeduDth	['bahkan', 'pas', 'di', 'bandung', 'pun', 'beli', 'oleh', 'oleh', 'via', 'gojek', 'kirim', 'ke', 'hotel', 'karena', 'promo', 'gofood', 'gede', 'bangett', 'bisa', 'nyenengin', 'semua', 'orang', 'temen', 'kantor', 'ayah', 'temen', 'mama', 'tetangga', 'dengan', 'biaya', 'hemat', 'terimakasih', 'gojek', 'gofood', 'cepatangofood']

4) *Stopword*

Filtering atau *stopword* merupakan tahapan untuk menghilangkan kata yang tidak penting dalam sebuah kalimat [12]. Tahap *stopword* dapat dilihat pada tabel 5.:

Tabel 5. *Stopword*

Tweet	Stopword
Bahkan pas di Bandung pun beli oleh oleh via gojek kirim ke hotel karena promo gofood gede banget <input type="checkbox"/> bisa nyenengin semua orang temen kantor ayah temen mama tetangga dengan biaya hemat <input type="checkbox"/> makasih gojek gofood #CepetanGoFood https://t.co/DENBcdUth	['pas', 'bandung', 'beli', 'oleh', 'oleh', 'via', 'gojek', 'kirim', 'hotel', 'promo', 'gofood', 'gede', 'bangett', 'bisa', 'nyenengin', 'semua', 'orang', 'temen', 'kantor', 'ayah', 'temen', 'mama', 'tetangga', 'biaya', 'hemat', 'terimakasih', 'gojek', 'gofood', 'cepatangofood']

5) *Stemming*

Stemming adalah mengubah kata ke bentuk dasarnya, suatu proses yang mengurangi variasi kata dengan kata dasar yang sama. Misalnya pada kata “diperoleh”, setelah melalui proses *stemming* maka akan menjadi kata “oleh” [13]. Tahap *stemming* dapat dilihat pada tabel 6.:

Tabel 6. *Stemming*

Tweet	Stemming
Bahkan pas di Bandung pun beli oleh oleh via gojek kirim ke hotel karena promo gofood gede banget <input type="checkbox"/> bisa nyenengin semua orang temen kantor ayah temen mama tetangga dengan biaya hemat <input type="checkbox"/> makasih gojek gofood #CepetanGoFood https://t.co/DENBcdUth	['waktu', 'bandung', 'beli', 'oleh', 'oleh', 'via', 'gojek', 'kirim', 'hotel', 'promo', 'gofood', 'besar', 'bangett', 'bisa', 'senang', 'semua', 'orang', 'temen', 'kantor', 'ayah', 'temen', 'mama', 'tetangga', 'biaya', 'hemat', 'terimakasih', 'gojek', 'gofood', 'cepatangofood']

20

3. Pembobotan kata (TF-IDF)

Analisis frekuensi kata TF-IDF (*Term Frequency-Inverse Document Frequency*) digunakan untuk mengimplementasikan pembobotan kata yang bekerja dengan cara menghitung frekuensi setiap kata dan melaporkan frekuensi kata terpenting, serta frekuensi kemunculan kata dalam dokumen. Perhitungan TF-IDF dapat dilihat pada persamaan (2)

$$W_{dt} = TF_{dt} \times \log\left(\frac{d}{df}\right) \quad (2)$$

Keterangan :

TF = kemunculan kata pada dokumen

df = kemunculan kata pada kumpulan dokumen

d = jumlah dokumen

df = banyak dokumen yang mengandung kata (term)

4. Klasifikasi analisis sentimen

Setelah *dataset* di *preprocessing* dan kemudian dilakukan pembobotan kata, tahap selanjutnya yaitu mengklasifikasikan ke dalam metode *K-Nearest Neighbor* dengan cara mengambil hasil pembobotan kata. *Dataset* bersih akan dibagi dengan kombinasi 80% *data training* dan 20% *data testing*. Tentukan nilai *k*, Selanjutnya *data training* akan dilatih oleh sistem sehingga menghasilkan model K-NN, lalu model K-NN digunakan untuk memprediksi hasil klasifikasi menggunakan *data testing*.

5. Metode Evaluasi dengan *Confusion Matrix*

Confusion matrix digunakan sebagai evaluasi klasifikasi metode *K-Nearest Neighbor* (K-NN). *Confusion matrix* menggunakan *confusion matrix* diharapkan dapat mengetahui seberapa akurat metode yang di evaluasi [14]. Parameter yang digunakan yaitu *accuracy*, *precision*, *recall*, *f1-score*. Tabel 7. Adalah gambaran *confusion matrix*.

Tabel 7. *Confusion Matrix*

Kondisi	Nilai Sebenarnya	
	Terklasifikasi Positif	Terklasifikasi Negatif
Prediksi	True Positif (TP)	False Negatif (FN)
	False Positif (FP)	True Negatif (TN)

Keterangan:

True Positive (TP) = Jumlah prediksi yang benar dari data terklasifikasi positif

False Positive (FP) = Jumlah prediksi yang benar yang terklasifikasi negatif

True Negative (TN) = Jumlah prediksi yang salah dari data yang terklasifikasi positif

False negative (FN) = Jumlah prediksi yang salah dari data yang terklasifikasi negatif

Accuracy merupakan perhitungan untuk menghilangkan label prediksi yang sesuai pada label aktual [15]. Jika hasil akurasi yang didapatkan pada perhitungan lebih tinggi dari prediksinya, maka algoritma yang digunakan baik. Pengambilan nilai akurasi dari hasil penambahan true positif dan true negatif, yang jumlahnya akan dibagikan dengan semua jumlah data. Perhitungan *Accuracy* dapat dilihat pada persamaan (3)

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \times 100\% \quad (3)$$

Precision (ketepatan) merupakan perhitungan perbandingan data yang dikenal didalam sistem dengan jumlah dalam dokumen yang dikenal

sistem. Perhitungan ² *Precision* dapat dilihat pada persamaan (4)

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

Recall (keberhasilan) perhitungan perbandingan total data berkaitan dengan ² semua total data yang berkaitan. Perhitungan *Recall* dapat dilihat pada persamaan (5)

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

Nilai dari *f1-score* merupakan suatu nilai yang digunakan untuk mengetahui nilai kinerja dari algoritma *K-Nearest Neighbor*. Perhitungan *f1-score* dapat dilihat pada persamaan (6)

$$f1\text{-score} = 2 \times \frac{precision * recall}{precision + recall} \quad (6)$$

3. HASIL DAN PEMBAHASAN

a. Klasifikasi Algoritma K-Nearest Neighbor

Untuk penelitian ini, peneliti ²⁴ menggunakan 1409 data. Peneliti kemudian membagi *dataset* menjadi dua bagian *data training* dan *data testing*, dengan pembagian 80:20 antara dua bagian data yang awalnya dikumpulkan, dengan *query* yang tersedia untuk dilihat hasil akurasi dan evaluasi menggunakan *confusion matrix*. *Split* data dilakukan secara *random by machine*.

Klasifikasi algoritma *K-Nearest Neighbor* yang dilakukan dalam ³ analisa ini menggunakan perbandingan 8:2 yaitu 80% data *training* dan 20% data *testing*. Data sebanyak 1409 data dengan *query* yang tersedia untuk dilihat hasil akurasi dan evaluasi menggunakan *confusion matrix*. Pada proses ini sistem juga dapat mengunggah *dataset* kedalam sistem untuk dilakukan proses klasifikasi.

Klasifikasi menghasilkan evaluasi dari *confusion matrix* berupa *accuracy K-Nearest Neighbor*, *recall*, *f1-score*, dan *precision*. Dari hasil pengukuran menggunakan 1409 data yang telah diambil, dilakukan uji coba klasifikasi.

Tabel 7: Uji Coba Klasifikasi

K	Accuracy	Precision	Recall	f1-Score
1	0.762411	0.848738	0.795276	0.821138
2	0.765957	0.848739	0.795276	0.821138
3	0.755319	0.853448	0.779528	0.814815
4	0.773050	0.875000	0.771654	0.820084
5	0.776596	0.881818	0.763780	0.818565
6	0.776596	0.881818	0.763780	0.818565
7	0.790780	0.915094	0.763780	0.832618

K	Accuracy	Precision	Recall	f1-Score
8	0.787234	0.932039	0.755906	0.834783
9	0.790780	0.960000	0.755906	0.845815
10	0.780142	0.940594	0.748031	0.833333
11	0.783688	0.923077	0.755906	0.831169
12	0.776596	0.913462	0.748031	0.822511
13	0.794326	0.941176	0.755906	0.838428
14	0.787234	0.931373	0.748031	0.829694
15	0.794326	0.950495	0.755906	0.842105
16	0.787234	0.914286	0.755906	0.827586
17	0.780142	0.914286	0.755906	0.827586
18	0.783688	0.914286	0.755906	0.827586
19	0.787234	0.923077	0.755906	0.831169
20	0.780142	0.914286	0.755906	0.827586

Dari hasil pengujian nilai K dapat dilihat ² pada tabel 4.7, penelitian ini menggunakan k=1, k=2, k=3, k=4, k=5, k=6, k=7, k=8, k=9, k=10, k=11, k=12, k=13, k=14, k=15, k=16, k=17, k=18, k=19, k=20. Dapat disimpulkan bahwa K dengan hasil klasifikasi terbaik pada k=15 dengan didapatkan *accuracy* sebesar 79,43%. ³³ pengujian klasifikasi algoritma *K-Nearest Neighbor* dapat dilihat pada Gambar 1:

	precision	recall	f1-score	support
negatif	0.444444	0.75915	0.250000	25.000000
netral	0.720930	0.939304	0.85798	152.000000
positif	0.900495	0.759304	0.842105	127.000000
accuracy	0.794326	0.794326	0.794326	0.794326
macro_avg	0.782290	0.828071	0.809965	182.000000
weighted_avg	0.807364	0.794326	0.794995	282.000000

Gambar 1. Klasifikasi Algoritma K-Nearest Neighbor

b. Evaluasi dengan Confusion Matrix

Evaluasi dengan *Confusion matrix* terhadap 1409 data dengan menggunakan parameter yang digunakan yaitu *accuracy*, *precision*, *recall*, *f1-score*.

Accuracy merupakan nilai *predicted* dari hasil perhitungan *true positif* dan *true negative* dari data yang diambil. Hasil perhitungan dari akurasi adalah 79,43%. Adapun penjabaran perhitungan dari akurasi adalah sebagai berikut:

$$Accuracy = \frac{(96 + 124 + 4)}{(96 + 31 + 0 + 3 + 124 + 5 + 2 + 17 + 4)} \times 100\%$$

$$Accuracy = \frac{224}{282} \times 100\%$$

$$Accuracy = 79,43\%$$

Precision merupakan hasil perhitungan dari nilai *predicted true positif* dibagi dengan total nilai *predicted positif*. Hasil perhitungan dari *precision*

adalah positif 0.95, netral 0.72, dan negatif 0.44. Adapun penjabaran perhitungan dari precision adalah sebagai berikut:

$$\text{Precision positif} = \frac{96}{(96 + 3 + 2)} = \frac{96}{101}$$

Precision positif = 0,95

$$\text{Precision Netral} = \frac{124}{(31 + 124 + 17)} = \frac{124}{172}$$

Precision Netral = 0,72

$$\text{Precision Negatif} = \frac{4}{(0 + 5 + 4)} = \frac{4}{9}$$

Precision Negatif= 0,44

Recall merupakan hasil perhitungan dari nilai *predicted true* positif dibagi dengan total nilai positif. Hasil perhitungan dari *recall* adalah positif 0.76, netral 0.94, dan negatif 0.17. Adapun penjabaran perhitungan dari *recall* adalah sebagai berikut:

$$\text{Recall positif} = \frac{96}{(96 + 31 + 0)} = \frac{96}{127}$$

Recall positif= 0,76

$$\text{Recall netral} = \frac{124}{(3 + 124 + 5)} = \frac{124}{132}$$

Recall netral = 0,94

$$\text{Recall negatif} = \frac{4}{(2 + 17 + 4)} = \frac{4}{23}$$

Recall negatif = 0,17

37 *f1-score* merupakan rata-rata dari hasil perhitungan *precision* dan *recall*. Hasil perhitungan dari *recall* adalah positif 0.84, netral 0.81, dan negatif 0.25. Adapun penjabaran perhitungan dari *f1-score* adalah sebagai berikut:

$$f1\text{-score positif} = 2 \times \frac{0,95 \times 0,76}{0,95 + 0,76} = 2 \times \frac{0,72}{1,71}$$

f1-score positif = 0.84

$$f1\text{-score netral} = 2 \times \frac{0,72 \times 0,94}{0,72 + 0,94} = 2 \times \frac{0,67}{1,66}$$

f1-score netral = 0.81

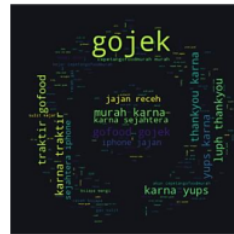
$$f1\text{-score negatif} = 2 \times \frac{0,44 \times 0,17}{0,44 + 0,17} = 2 \times \frac{0,074}{0,61}$$

f1-score negatif = 0.25

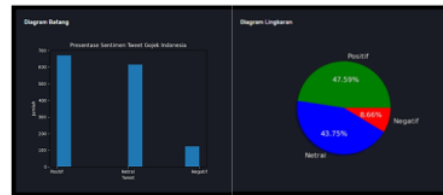
Hal yang paling mempengaruhi dalam proses klasifikasi yaitu ketika *tweet* yang didapatkan masih banyak *tweet* yang menggunakan kata singkatan, hal itu akan sangat berpengaruh pada nilai prediksi yang didapatkan dan berpengaruh juga pada tingkat akurasi.

c. Visualisasi

Visualisasi ditampilkan sebagai bentuk persentase hasil klasifikasi, berupa diagram batang, *pie chart*, serta *wordcloud*. Tujuan peneliti menampilkan visualisasi untuk mempermudah melihat visual p40rma metode yang digunakan. Hasil Visualisasi dapat dilihat pada gambar dibawah ini.



Gambar 3. Wordcloud



Gambar 4. Visualisasi Diagram Batang dan pie chart

28

4. Kesimpulan dan Saran

a. Kesimpulan

Berdasarkan hasil penelitian dengan **39** menggunakan 1.409 data menggunakan metode *K-Nearest Neighbor* mendapatkan akurasi sebesar 79.43% pada k=15, dengan sentimen positif sebesar 47.59%, negatif 8.66%, dan netral 43.75%. Penelitian ini diawali dengan *crawling* data, selanjutnya data mentah diolah terlebih dahulu pada proses *preprocessing*. Setelah itu data bersih dilakukan *labellit* **36** (pembobotan kata), kemudian tahap klasifikasi menggunakan metode *K-Nearest Neighbor* dengan perhitungan jarak *euclidean*. Tahap akhir yaitu tahap evaluasi terhadap metode yang digunakan dengan *confusion matrix* dengan *split* data 80:20 data *training* dan *data testing*.

b. Saran

1. Meningkatkan kosakata baku dan tidak baku agar data semakin akurat dan stabil mengingat *dataset* banyak menggunakan kata tidak baku, singkatan dan bahasa gaul lainnya.
2. Penelitian hanya dilakukan pada satu metode saja, untuk kedepannya bisa dilakukan dengan membandingkan dua metode atau lebih.
3. Pada penelitian ini hanya fokus pada satu bahasa saja yaitu Bahasa Indonesia. Kedepannya bisa dikembangkan lagi dengan bahasa Inggris

mengingat gojek sudah tersebar luas di banyak negara.

DAFTAR PUSTAKA

- [1] D. A. P. Wulandari, M. Sudarma, and N. Paramaita. 2019. "Pemanfaatan Big Data Media Sosial Dalam Menganalisa Kemenangan Pilkada," *Maj. Ilm. Teknol. Elektro*, vol. 18, no. 1, p. 101, doi: 10.24843/mite.2019.v18i01.p15.
- [2] A. Nurzahputra and M. A. Muslim. 2016. "Analisis Sentimen pada Opini Mahasiswa Menggunakan Natural Language Processing," *Semin. Nas. Ilmu Komput.*, no. Snik, pp. 104–118.
- [3] M. M. Yulian Findawati, S.T. and M. K. Muhammad Alfian Rosid, S.K. 2020. *BUKU AJAR TEXT MINING*. UMSIDA Press Anggota IKAPI No. 218/Anggota Luar Biasa/JTI/2019 Anggota APPTI No. 002 018 1 09 2017.
- [4] H. S. Jose, P. A. Firdaus, M. Khalifah, and N. Ayuningsih. 2021. "CORPORATE DIPLOMACY GO-JEK MELALUI PENDIRIAN GO-VIET 2018-2020: TINJAUAN NATION BRANDING INDONESIA DI VIETNAM," vol. 6, no. 9, pp. 817–828.
- [5] M. M. Baharuddin, H. Azis, and T. Hasanuddin. 2019. "Analisis Performa Metode K-Nearest Neighbor Untuk Identifikasi Jenis Kaca," *Ilk. J. Ilm.*, vol. 11, no. 3, pp. 269–274, doi: 10.33096/ilkom.v11i3.489.269-274.
- [6] E. R. Mudy, G. Raharjito, A. Susanto, and M. Muslih. 2020. "Implementasi Algoritma K-Neares Neighbor dalam Pengenalan Aksara Jawa Berbasis Android untuk Anak Sekolah Dasar," *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 2, no. 2, p. 7, doi: 10.36499/jinrpl.v2i2.3543.
- [7] R. D. Liklikwatil, E. Noersongko, and C. Supriyanto. 2018. "Optimasi K-Nearest Neighbor Dengan Particle Swarm Optimization Untuk Memprediksi Harga Komoditi Karet," *e-Jurnal JUSITI (Jurnal Sist. Inf. dan Teknol. Informasi)*, vol. 7–2, no. 2, pp. 172–182, doi: 10.36774/jusiti.v7i2.252.
- [8] Khairul Hudha Nasution, Widodo, and Bambang Prasetya Adhi. 2021. "Sistem Deteksi Topik Politik Pada Twitter Menggunakan Algoritma Latent Dirichlet Allocation," *PINTER J. Pendidik. Tek. Inform. dan Komput.*, vol. 6, no. 1, pp. 76–83, doi: 10.21009/pinter.5.1.10.
- [9] M. W. A. Putra, Susanti, Erlin, and Herwin. 2020. "Analisis Sentimen Dompot Elektronik Pada Twitter Menggunakan Metode Naïve Bayes Classifier," *IT J. Res. Dev.*, vol. 5, no. 1, pp. 72–86, doi: 10.25299/itjrd.2020.vol5(1).5159.
- [10] S. Hikmawan, A. Pardamean, and S. N. Khasanah. 2020. "Sentimen Analisis Publik Terhadap Joko Widodo terhadap wabah Covid-19 menggunakan Metode Machine Learning," *J. Kaji. Ilm.*, vol. 20, no. 2, pp. 167–176, doi: 10.31599/jki.v20i2.117.
- [11] D. S. V. and M. R. Janani. 2016. "Text Mining: open Source Tokenization Tools – An Analysis," *Adv. Comput. Intell. An Int. J.*, vol. 3, no. 1, pp. 37–47, doi: 10.5121/acii.2016.3104.
- [12] N. Fitriyah, B. Warsito, and D. A. I. Maruddani. 2020. "Analisis Sentimen Goejek Pada Media Sosial Twitter Dengan Klasifikasi Support Vector Machine (Svm)," *J. Gaussian*, vol. 9, no. 3, pp. 376–390, doi: 10.14719/j.gauss.v9i3.28932.
- [13] A. A. Irfa, Adiwijaya, and M. S. Mubarak. 2018. "Klasifikasi Topik Berita Berbahasa Indonesia Menggunakan k-Nearest Neighbor," *e-Proceeding Eng.*, vol. 5, no. 2, pp. 631.
- [14] H. S. Raja and S. Magdhalena. 2019. "Twitter Sentimen Goejek Indonesia Dan Grab," *Pros. Semin. Nas. Mat. Stat. dan Apl.*, vol. 1, pp. 256–261.
- [15] M. Papuungan. 2018. "Penerapan Case Based Reasoning Untuk Sistem Diagnosis Penyakit Hepatitis," *JIKO (Jurnal Inform. dan Komputer)*, vol. 1, no. 1, pp. 7–12, doi: 10.33387/jiko.v1i1.1165.

ORIGINALITY REPORT

21 %
SIMILARITY INDEX

17 %
INTERNET SOURCES

10 %
PUBLICATIONS

7 %
STUDENT PAPERS

PRIMARY SOURCES

1 Submitted to Universitas Brawijaya **2** %
Student Paper

2 repository.ub.ac.id **2** %
Internet Source

3 123dok.com **1** %
Internet Source

4 Andhika Ryan Pratama, Muhammad Mustajib,
Aryo Nugroho. "Deteksi Citra Uang Kertas
dengan Fitur RGB Menggunakan K-Nearest
Neighbor", Jurnal Eksplora Informatika, 2020 **1** %
Publication

5 jurnal.iaii.or.id **1** %
Internet Source

6 www.ejurnal.stmik-budidarma.ac.id **1** %
Internet Source

7 www.trilogi.ac.id **1** %
Internet Source

8 repository.unja.ac.id **1** %
Internet Source

9	ejurnal.stmik-budidarma.ac.id Internet Source	1 %
10	press.umsida.ac.id Internet Source	1 %
11	Kevin Djajadinata, Hussein Faisol, Guruh Fajar Shidik, Muljono, Ahmad Zainul Fanani. "Evaluation of Feature Extraction for Indonesian News Classification", 2020 International Seminar on Application for Technology of Information and Communication (iSemantic), 2020 Publication	1 %
12	ojs.unpkediri.ac.id Internet Source	1 %
13	www.coursehero.com Internet Source	<1 %
14	sismatik.nusaputra.ac.id Internet Source	<1 %
15	www.jurnal.iaii.or.id Internet Source	<1 %
16	Submitted to President University Student Paper	<1 %
17	inacl.id Internet Source	<1 %
18	ojs.unikom.ac.id	

Internet Source

<1 %

19

prosiding.senadi.upy.ac.id

Internet Source

<1 %

20

Submitted to Universitas Jember

Student Paper

<1 %

21

www.unisbank.ac.id

Internet Source

<1 %

22

yuita.lecture.ub.ac.id

Internet Source

<1 %

23

www.scribd.com

Internet Source

<1 %

24

Submitted to Universitas Budi Luhur

Student Paper

<1 %

25

jurnal.stts.edu

Internet Source

<1 %

26

adi-journal.org

Internet Source

<1 %

27

www.researchgate.net

Internet Source

<1 %

28

digilib.unila.ac.id

Internet Source

<1 %

29

conference.upnvj.ac.id

Internet Source

<1 %

30	ejournal.unkhair.ac.id Internet Source	<1 %
31	ojs.unud.ac.id Internet Source	<1 %
32	ttkhok.elte.hu Internet Source	<1 %
33	www.informatika.org Internet Source	<1 %
34	www.jurnal.stkipppgritulungagung.ac.id Internet Source	<1 %
35	Luluk Suryani, Kasmi Edy. "PENGEMBANGAN APLIKASI "LOST & FOUND" BERBASIS ANDROID DENGAN MENGGUNAKAN METODE TERM FREQUENCY – INVERSE DOCUMENT FREQUENCY (TF-IDF) DAN COSINE SIMILARITY", Electro Luceat, 2020 Publication	<1 %
36	Rafika Amilia, Eko Prasetyo. "KLASIFIKASI DIAGNOSA PENYAKIT DEMAM BERDARAH DENGUE PADA ANAK MENGGUNAKAN METODE K-NEAREST NEIGHBOR STUDI KASUS RUMAH SAKIT PKU MUHAMMADIYAH UJUNG PANGKAH GRESIK", Indexia, 2021 Publication	<1 %
37	Santi Thomas, Yuliana, Noviyanti. P. "Study Analisis Metode Analisis Sentimen pada	<1 %

YouTube", Journal of Information Technology, 2021

Publication

38

Toni Arifin. "IMPLEMENTASI METODE K-NEAREST NEIGHBOR UNTUK KLASIFIKASI CITRA SEL PAP SMEAR MENGGUNAKAN ANALISIS TEKSTUR NUKLEUS", Jurnal Informatika, 2016

Publication

<1 %

39

Zulfrianto Y Lamasigi, Maryam Hasan, Yulianti Lasena. "Local Binary Pattern untuk Pengenalan Jenis Daun Tanaman Obat menggunakan K-Nearest Neighbor", ILKOM Jurnal Ilmiah, 2020

Publication

<1 %

40

catatankuliahku2008.wordpress.com

Internet Source

<1 %

41

etheses.uin-malang.ac.id

Internet Source

<1 %

42

journal.ikopin.ac.id

Internet Source

<1 %

43

pdfs.semanticscholar.org

Internet Source

<1 %

44

repository.uin-suska.ac.id

Internet Source

<1 %

45

repository.usd.ac.id

Internet Source

<1 %

46

Hani - Harafani. "Forward Selection pada Support Vector Machine untuk Memprediksi Kanker Payudara", Jurnal Infortech, 2020

Publication

<1 %

Exclude quotes On

Exclude matches Off

Exclude bibliography Off