Jurnal Ilmiah Informatika (JIF)
| ISSN (Print) 2337-8379 | ISSN (Online) 2615-1049 |
Jurnal online di akses di http://ejournal.upbatam.ac.id/index.php/jif

Perbandingan Algoritma Naïve Bayes Classifier Dan K-Nearest Neighbors Untuk Analisis Sentimen Covid-19 Di Twitter

Habibi Aulia Nur Syifa[1], Arie Nugroho[2], Rina Firliana[3]

[1,2,3]Sistem Informasi Fakultas Teknik, Universitas Nusantara PGRI Kediri, Kampus II. Mojoroto Gang I No.6 Kediri 64112, Indonesia

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 28 Januari 2023
Revisi Akhir: 16 Februari 2023
Diterbitkan *Online*: 10 Maret 2023

KATA KUNCI

Covid-19
Twitter
Analisis Sentimen
Naïve Bayes Classifier
K-Nearest Neighbors

KORESPONDENSI

E-mail: habibiaulianursyifa@gmail.com

A B S T R A C T

COVID-19 emerged in China in 2019. In Indonesia in 2020 there were more than 3000 positive cases of COVID-19 with a mortality rate of 9.1%. The government's incomplete efforts to break the chain of the spread of COVID-19 have made people uneasy about this pandemic. Many people want to express their aspirations on social media which is considered suitable as a place that represents the aspirations of the COVID-19 pandemic. One of them is twitter. There are so many text messages sent, some positive and some negative, that it is difficult to retrieve harmonized information due to the diversity of text messages sent. One way to overcome this is with sentiment analysis. This research has processes including text preprocessing, word weighting, classification with K-Nearest Neighbors (KNN) and Naïve Bayes Classifiers (NBC) algorithms. The results obtained by KNN got an accuracy of 72.37% while NBC amounted to 67.84%. KNN is the best classification algorithm for negative sentiment classification, the negative label predicted correctly in KNN is greater, namely 393 compared to NBC which is 339. while NBC is the best algorithm for positive sentiment classification, the positive label predicted correctly NBC is 275 greater than KNN as much as 262.

1. PENDAHULUAN

Banyak area komunikasi yang telah berubah sebagai hasil dari ekspansi internet yang cepat. Semakin banyak platform komunikasi termasuk media sosial yang muncul sebagai hasil dari ekspansi internet. Di Indonesia pengguna media sosial mencapai 170 juta pengguna dengan 5 teratas yang paling populer yakni *youtube*, *whatsapp*, *instagram*, *facebook* dan *twitter* [1].

Salah satu platform media sosial yang paling banyak digunakan untuk komunikasi sosial ialah *Twitter*. Pengguna *Twitter* dapat mengirim dan menerima pesan secara langsung kepada pengguna lain. Di Indonesia *Twitter* mendapatkan peringkat ke-5 sebagai media sosial yang paling banyak digunakan Menjadikan *Twitter* sebagai salah satu *platform* untuk beropini dan menyampaikan pendapat masyarakat mengenai berbagai isu dan perbincangan selama wabah COVID-19 di Indonesia [2]

COVID-19 atau dapat disebut C*oronavirus 19* ialah infeksi gangguan pernapasan yang dapat menular dengan cepat. Penyakit COVID-19 muncul di daerah Wuhan, Tiongkok pada bulan November 2019 [3]. *World Health Organization* selanjutnya memberikan penamaan jenis corona virus baru ini sejak terjadinya peristiwa menggemparkan di Tiongkok pada Desember 2019. SARS-COV2 ialah virus penyebab COVID-19. Sejak kedatangannya di Indonesia pada awal maret lalu, Virus COVID-19 diperkirakan masih terus berkembang[4].

Di negara Indonesia sendiri pada tanggal 10 april 2020 ditemukan sebanyak lebih dari 3000 kasus positif mengidap *COVID-19*, 282 orang sembuh dan 306 orang meninggal dunia dengan tingkat kematian sebesar 9.1% [5]. Masyarakat merasa resah dan takut akan pandemi ini karena upaya pemerintah yang tidak efektif dalam memutus rantai penularan COVID-19 [6]. Banyak masyakarat yang ingin mengekspresikan aspirasinya di media sosial yang tampaknya tepat sebagai tempat untuk mengungkapkan aspirasi masyakarat akan pandemi *COVID-19*. Salah satu media sosialnya yaitu *twitter*, dimana pada *twitter* dimungkinkan pengguna akun mengirimkan hingga 140 karakter pada pesan teksnya [7]. Terdapat banyak sekali pesan teks yang dikirimkan ada pesan teks yang positif dan ada juga pesan teks yang negatif sehingga sulit untuk mengambil informasi yang selaras didalamnya karena keberagaman pesan teks yang dikirim. Menerapkan analisis sentimen(pendapat) ialah salah satu metode untuk mengatasi hal ini.

Analisis sentimen(pendapat) masuk kedalam bidang diantara beberapa bidang penelitian seperti pemrosesan bahasa alami, pembelajaran mesin dan penambangan data. Analisis sentimen atau opini merupakan pemisahan sentimen dari isi sebuah kalimat atau teks. Analisis sentimen(pendapat) dilakukan untuk mengelompokkan kalimat atau sentimen berdasarkan pola sifat didalamnya. Pola sifat ini merupakan sentimen yang memiliki aspek positif, negatif atau netral [8]. Penelitian sebelumnya perihal analisis sentimen di media sosial *twitter* terkait dengan COVID-19 [7]. Penelitian tersebut menggunakan kamus sastrawi dalam preprocessingnya, menggunakan algoritma *Naïve Bayes Classifiers* dan mendapatkan hasil akurasi sebesar 78% .Penelitian lain terkait dengan COVID-19 [9]. Penelitian tersebut menggunakan data private berjumlah 796 *tweet* serta algoritma *Naïve Bayes Classifiers* dan mendapat akurasi sebesar 67%.[8] dalam penelitiannya menggunakan 3 tahap *preprocessing* yaitu *case folding, tokenize dan stopword*, menggunakan algoritma *naïve bayes classifiers* dan didapat hasil akurasi sebesar 69%.

Keunikan penelitian ini dibandingkan dengan penelitian sebelumnya adalah penerapan tahap *preprocessing* yang lebih banyak serta pembuatan kamus *stopword* dan *stemming* sendiri yang didasarkan pada dataset yang dimiliki serta penggunaan data publik. Penulis menggunakan *rapidminer* sebagai alat bantu menganalisis data. Dalam pengumpulan datanya penulis menggunakan dataset publik yang diambil dari website *kaggle* yang memiliki pesan teks bersentimen negatif dan positif [10]. Dataset kemudian dilakukan pemberian nilai bobot pada tiap kata untuk mengetahui dampak sebuah kata dalam dataset. Teknik yang dipakai ialah *Term Frequency-Inverse Document Frequency* (TF-IDF), teknik pemberian nilai bobot kata dari sebuah dataset menggunakan *transform cases*, tokenisasi, *normalization*, *stopwords*, dan *stemming*. Metode ini digunakan karena mudah dipahami dan diterapkan pada permasalahan keakuratan dokumen[11]. Data yang telah diproses menggunakan TF-IDF kemudian dilakukan pengklasifikan untuk mendapat sentimen positif dan negatif. Metode yang dipakai yaitu *Naïve Bayes Classifier* (NBC) ditambah *K-Nearest Neighbors* (KNN) sebagai perbandingan metode. Karena mudah dalam melatih dan serta menghasilkan akurasi yang tinggi ketika mengklasifikasikan data yang berjumlah banyak, NBC sering dipakai dalam pendekatan klasifikasi [12].

Maka dari itu, berdasarkan uraian diatas penulis mengusulkan untuk membuat sebuah penelitian berjudul Perbandingan Algoritma *Naïve Bayes Classifier* Dan *K-Nearest Neighbors* Untuk Analisis Sentimen *Covid-19* Di *Twitter* dan diharapkan penelitian ini dapat membantu dalam mengetahui kecenderungan opini masyarakat terkait *COVID-19*

2. TINJAUAN PUSTAKA

2.1 *Sentiment Analysis*

Analisis Sentimen ialah proses menemukan sentimen dari sebuah teks atau kalimat apakah positif, negatif atau netral [13]. Pendekatan ini juga dipakai untuk mengetahui pendapat publik terhadap desas-desus, topik, kebijakan dan pelayanan berdasarkan data tekstual [14] . Analisis sentimen memiliki beberapa tahap antara lain: [15].

a. Pengumpulan data. Yaitu prosedur untuk mengumpulkan data yang akan dianalisis.
b. Preprocessing, proses untuk mendapatkan data yang sesuai untuk klasifikasi. Langkah yang dipakai untuk menghapus data dari noise, menselaraskan bentuk kata dan mengurangi ukuran data untuk memperoleh data yang paling cocok dipakai saat pengklasifikasian.
c. Klasifikasi, ialah proses untuk mengetahui pola sentimen(pendapat) dari pesan teks yang telah melalui proses prerpocessing dimana dataset pesan teks diterapkan menggunakan algoritma klasifikasi.
d. Evaluasi performa, yaitu proses untuk menghitunga akurasi, presisi, recall dan f-measure.

2.2 *Preprocessing*

Preprocessing ialah proses untuk memperoleh data yang cocok untuk dipakai saat dan memiliki beberapa langkah yang dilakukan antara lain: [16].

a. *Cleansing* atau proses untuk menghapus username, hashtag pada pesan teks.
b. *Case folding*, yaitu proses untuk menghubah huruf pada kalimat di dalam pesan teks menjadi huruf kecil.
c. Normalisasi, yaitu proses merubah kata yang tidak tepat menjadi kata yang tepat mengacu pada Kamus Besar Bahasa Indonesia (KBBI).
d. *Tokenization*, ialah proses untuk mengubah dokumen atau kalimat pada pesan teks menjadi kumpulan token kata.
e. *Stopword Removal*, ialah langkah untuk menghapus kata yang tidak perlu atau kurang bermakna.
f. *Stemming*, yaitu langkah merubah kata berimbuhan ke dalam bentuk kata dasarnya.

2.3 Term-Frequency-Inverse Document Frequency

Teknik perhitungan atau pemberian nilai bobot tiap kata dalam dokumen yang berfungsi untuk menunjukan pentingnya kata dalam dokumen tersebut dengan sebelumnya melakukan beberapa proses seperti tokeninasi, *stopword*, *stemming* dan frekuensi kemunculan kata dalam dokumen [11] . Rumus *TF-IDF* dapat dilihat sebagai berikut [17].

$$W_{t,d} = tf_{t,d} \times idf_t = tf_{t,d} \times \log \frac{N}{df_t} \qquad (1)$$

Dimana (1):

$W_{t,d}$ = Nilai TF-IDF

$tf_{t,d}$ = Nilai kemunculan kata (t) pada dokumen (d)

idf_t = Nilai inverse dokumen settiap kata (t)

df_t = Nilai kemunculan dokumen setiap kata(t)

N = Nilai seluruh dokumen

2.4 Naïve Bayes Classifiers

Naïve Bayes Classifiers (NBC) atau *Bayesian Classifiers* ialah metode yang digunakan untuk memprediksi suatu kemungkinan untuk memutuskan kategori kelas dokumen teks tertentu pada baris tertentu dan juga dapat menghasilkan akurasi yang cukup tinggi [18]..

2.5 K-Nearest Neighbors

Dalam [19] menyatakan bahwa algoritma KNN termasuk algoritma supervised learning atau algoritma yang menghubungkan pola data saat ini dengan data baru dalam upaya untuk menemukan pola baru dalam data. Untuk nilai prediksi data uji baru, KNN menggunakan klasifikasi tetangga.

2.6 Eucludian Distance

Eucludian Distance ialah metode jarak pengenalan pola yang sering dipakai untuk menentukan kesamaan antara dua pola, eucludian distance paling umum digunakan untuk melakukan perhitungan data numerik [19] . dimana eucludian distance memiliki rumus sebagai berikut [17].

$$d(A,B) = \sqrt{\sum_{k=1}^{n} (Xk - Xk)^2} \qquad (2)$$

Dimana (2):

D = jarak kemiripan antara dua titik A dan B

A = data yang diuji

B = sampel data

n = jumlah data

2.7 K-Fold Cross Validation

Sebuah teknik pengujian data dengan membagi dataset sejumlah n-fold kemudian dataset akan dibagi sejumlah n partisi dengan jumlah yang sama contoh partisi K, L, M, N dengan masing-masing partisi memiliki data yang berbeda kemudian akan di uji sejumlah n yang dimasukkan. Dalam contoh percobaan ke n partisi K akan dijadikan dataset uji dan sisa partisi lainnya akan menjadi dataset [20]. contoh skenario pengujian sebagai berikut

Tabel 1 Contoh Skenario Cross Validation

n-data				
	K	**L**	**M**	**N**
1	K	L	M	N
2	K	L	M	N
3	K	L	M	N
4	K	L	M	N
Data uji				
Data latih				

2.8 Confusion Matrix

Teknik yang berfungsi untuk mengevaluasi keampuhan metode klasifikasi dalam memprediksi kelas data dimana teknik ini memberikan perbandingan nilai kelas aslinya dengan nilai prediksi. Contoh tabel dapat dilihat di bawah ini. Pada *confusion matrix* dilakukan beberapa perhitungan sebagai berikut [21].

a. *Accurasy*, nilai dari data uji kelas diprediksi akurat oleh model klasifikasi selaras dengan kelas asli dengan menggunakan persamaan sebagai berikut.

$$accurasy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (3)$$

b. Presisi, ialah rasio kesamaan antara kelas data hasil prediksi algoritma dengan kelas data aslinya dengan menggunakan contoh persamaan berikut.

$$presisi = \frac{TP}{TP + FP} \qquad (4)$$

c. *Recall*, ialah nilai rasio dari ketepatan banyaknya data yang berhasil diprediksi oleh model algoritma berdasarkan kelas data aslinya menggunakan contoh persamaan berikut.

$$recall = \frac{TP}{TP + FN} \qquad (5)$$

d. *F-measure*, ialah perhitungan gabungan antara presisi dan *recall* dengan menggunakan contoh persamaan berikut.

$$F - measure = \frac{2xpresisixrecall}{presisi + recall} \qquad (6)$$

Tabel 2 Contoh Tabel Confusion Matrix

		Skor Terkini	
		(Positive)	(Negative)
Skor Perkiraan	(Positive)	TP	FP
	(Negative)	FN	TN

Pada tabel 2 confusion matrix terdapat empat nilai yang dikeluarkan antara lain [12] :

a. *True Positive* (TP), ialah nilai terkini yang bernilai positif dan terprediksi benar.
b. *True Negative* (TN), ialah nilai terkini yang bernilai negatif dan terprediksi benar
c. *False Positive* (FP), ialah nilai terkini yang bernilai negatif dan terprediksi positif
d. *False Negative* (FN), ialah nilai terkini yang bernilai positif dan terprediksi negatif.

2.9 Rapidminer

Rapidminer ialah sebuah platform perangkat lunak open source yang memiliki ekosistem kerja untuk pembelajaran mesin, pembelajaran mendalam, dan *predictive analysis*. Perangkat luan

ini dapat dipakai untuk bisnis, pendidikan, penelitian serta menunjang fase pembelajaran mesin seperti mempersiapkan data, visualisasi dan optimalisasi [22].

3. METODOLOGI

3.1. Metode Penelitian

Penelitian kausal komparatif atau penelitian perbandingan ialah Penelitian ini berfokus pada hubungan sebab-akibat dengan mencari tahu penyebab timbulnya perbedaan serta mengetahui apa yang berbeda, penelitian kausal komparatif dikategorikan sebagai penelitian kuantitatif [23]. Peneliti menggunakan penelitian komparatif karena merupakan penelitian yang membandingkan objek satu dengan objek yang lain. Peneliti ingin mencari tahu perbedaan antara metode NBC dan KNN serta menentukan metode mana yang terbaik untuk klasifikasi dataset *COVID-19* yang diteliti apakah ada perbedaan yang signifikan dari kedua metode tersebut

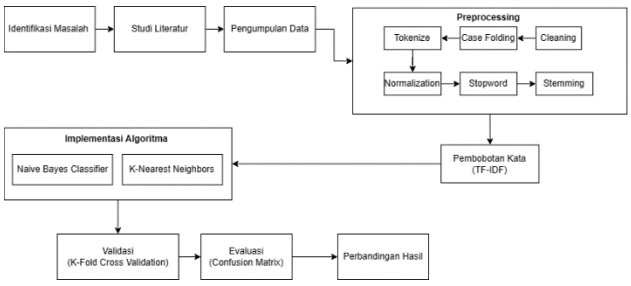
3.2. Pengumpulan Data

Penelitian menggunakan data *sekunder* dimana data diambil dari situs *kaggle* dan dataset berupa data publik mengenai komentar pada *twitter* terkait *COVID-19* [10].

3.3. Analisis Data

Penelitian menggunakan metode analisis deskrptif. Sebuah analisis statistik yang digunakan untuk menganalisis data dengan cara menggambarkan data apa adanya [24]. Analisis disajikan dalam bentuk grafik dan tabel. Alat yang dipakai untuk analisis data ialah rapidminer. Dimana dilakukan proses preprocessing, pembobotan kata dan pengklasifikasian naïve bayes dan KNN serta pembentukan grafik dan tabel

3.4. Alur Penelitian



Gambar 1 Alur Penelitian

Keterangan Gambar 1:

1. Pengumpulan data

Langkah untuk mempersiapkan data. Data yang digunakan ialah data sekunder, data diambil dari situs *kaggle* serta dataset berupa data publik mengenai komentar pada twitter terkait *COVID-19*.

2. *Preprocessing*

Pada tahap *pre-processing* ini data akan diolah secara bertahap agar menjadikan data lebih mudah diproses oleh sistem. *Preprocessing* dilakukan dengan menggunakan bantuan tools *rapidminer* Selanjutnya tahap yang dilakukan dalam *pre-processing* yaitu:

a. *Cleaning*, ini merupakan proses penghapusan kata yang tidak memiliki pengaruh seperti *url* dan *hastag*, *username* serta data duplikat.
b. *Case Folding* yaitu merubah huruf pada kalimat menjadi huruf kecil.
c. *Tokenization* langkah memisah kalimat menjadi kumpulan kata-kata. Dimana setiap kalimat dipecah menjadi kata per kata. Serta menghilangkan simbol dan angka
d. *Normalization*, yaitu proses menyesuaikan kata yang tidak sesuai atau tidak baku menjadi kata yang baku berdasarkan KBBI seperti kata singkatan.
e. *Stopword*, yaitu proses untuk menghilangkan kata yang tidak mengandung makna seperti kata sambung. Proses stopword menggunakan kamus stopword yang didapat dari hasil penelitian [25] dan kamus stopword yang telah dibuat berdasarkan temuan kata pada dataset yang dimiliki yang akan digambungkan menjadi satu file.
f. *Stemming*, langkah merubah kata berimbuhan menjadi bentuk dasarnya. Dilakukan menggunakan kamus yang dibuat berdasarkan temuan kata pada dataset dan berdasarkan KBBI.

3. Pembobotan Kata

Tahap dimana dilakukan Pembobotan setiap token kata pada data yang dipakai untuk menunjukkan pentingnya sebuah kata di dalam sebuah dokumen serta kata menunjukkan yang paling sering muncul. Pembobotan dilakukan menggunakan *TF-IDF* Dokumen yang digunakan dalam pembobotan kata ialah dataset hasil dari preproccesing.

4. Implementasi Algoritma

Langkah dilakukannya implementasi algoritma NBC dan KNN pada dataset *COVID-19* yang telah sebelumnya dilakukan preprocessing dan pembobotan kata.

5. Validasi

Langkah dilakukannya pengujian dengan metode *K-Fold Cross Validation*. Dataset hasil tahap implementasi algoritma dilakukan pengujian sebanyak 10 kali

6. Evaluasi

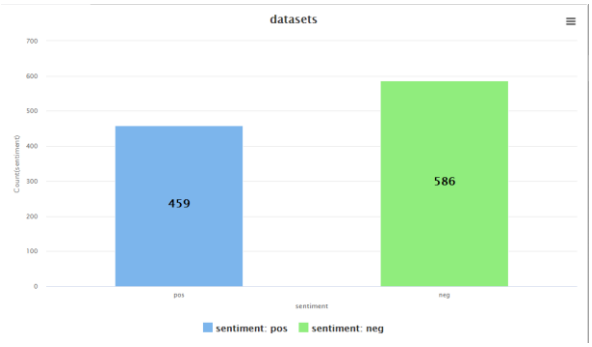
Tahap ini dilakukan evaluasi performa dari implementasi algoritma NBC dan KNN pada dataset *COVID-19* dengan menggunakan *confusion matrix*.

7. Perbandingan Hasil

Pada tahap ini hasil dari evaluasi yaitu tabel *confusion matrix* dari masing-masing akan dilakukan perbandingan dengan membandingkan hasil tabel *confusion matrix* antara algoritma NBC dan KNN

4. HASIL DAN PEMBAHASAN

4.1. Pengumpulan Data



Gambar 2 Proporsi Dataset

Dataset yang digunakan ialah data sekunder, data diambil dari situs *kaggle* serta data berupa data publik mengenai komentar pada *twitter* terkait *COVID-19*. Dengan jumlah data sebanyak 1045 pesan teks dengan 2 kategori label sentimen yaitu positif dan negatif dimana sebanyak 459 pesan teks berlabel positif dan 586 pesan teks berlabel negatif. terdapat 2 atribut yaitu *text* dan *sentiment* dan dataset ini akan digunakan untuk tahap *preprocessing*.

Tabel 3 Contoh Dataset

No	Text	Sentiment
1	Yukk sama-sama bahu membahu membantuu pemerintah memutus rantaii covid-19 dan memutus rantai kebenciann. Supaya Indonesia normal kembali	pos
2	Yuks.. kawal kebijakan pemerintah jangan sampai disalah gunakan oknum2 tidak bertanggung jawab. Indonesia Menang Lawan Covid-19 https://twitter.com/Ardhiona91/status/1245206909843427328Ã*ƒ*â€šÃ‚Â ÃƒÂ¢Ã¢â€šÂ¬Ã‚Â¦	pos
3	Yg perlu disalahkan bukan kang angkotnya, tapi pemerintah dan menkes, mereka tidak memberikan edukasi yang baik pada masyarakat tentang bahaya covid-19 dan penanggulangannya. Jangan salahkan mereka ngeyel, itu karma karena pemerintah ngeyel bhw Indonesia bebas virus.	neg
…	….	…
1045	Mengerikan menonton investigasi @narasitv tentang pasien covid-19 ke 0 di Indonesia. Betapa serampangannya pemerintah menangani hal seberbahaya ini. https://youtu.be/WJpBqfwDHsAÃ*ƒ*â€š	neg

4.2. *Preprocessing*

Sebelum data dilakukan proses klasifikasi, data diharuskan untuk diolah terlebih dahulu agar saat proses klasifikasi didapatkan hasil yang baik. Pengolahan data dilakukan dengan cara menyiapkan data sebelum diproses atau *pre-proccesing* serta menghapus data duplikat Langkah pengolahan dilakukan dengan tahapan dalam text mining. Langkah proses ini menggunakan operator yang terdapat pada ekstensi text processing di rapidminer.

4.2.1. Cleaning

pada proses ini pesan teks pada dataset dibersihkan dari teks yang tidak diinginkan seperti url atau link, username atau mentions, hashtag.

Tabel 4 Contoh Cleaning

Sebelum
Yuks.. kawal kebijakan pemerintah jangan sampai disalah gunakan oknum2 tidak bertanggung jawab. Indonesia Menang Lawan Covid-19 https://twitter.com/Ardhiona91/status/1245206909843427328Ã*ƒ*â€šÃ‚Â ÃƒÂ¢Ã¢â€šÂ¬Ã‚Â¦
Sesudah
Yuks.. kawal kebijakan pemerintah jangan sampai disalah gunakan oknum2 tidak bertanggung jawab. Indonesia Menang Lawan Covid-19

4.2.2. Case Folding

Pada proses ini dataset yang telah dilakukan cleaning sebelumnya akan dilakukan langkah untuk merubah semua huruf menjadi huruf kecil (*lower case*).

Tabel 5 Contoh Case Folding

Sebelum
Yuk Pak sampai kapan kita melihat kenaikan kasus positif Covid-19 hari ke hari, jumlah kematian yg terus meningkat. Mari cari solusi untuk membantu segala dampak buruk lockdown bagi masyarakat. Sampai kapan begini Pak??
Sesudah
yuk pak sampai kapan kita melihat kenaikan kasus positif covid-19 hari ke hari, jumlah kematian yg terus meningkat. mari cari solusi untuk membantu segala dampak buruk lockdown bagi masyarakat. sampai kapan begini pak??

4.2.3. Tokenize

Kemudian dilanjut dengan proses tokenize atau memecah kalimat menjadi kata per kata serta menghilangkan simbol dan angka.

Tabel 6 Contoh Tokenize

Sebelum				
yuk pak sampai kapan kita melihat kenaikan kasus positif covid-19 hari ke hari, jumlah kematian yg terus meningkat. mari cari solusi untuk membantu segala dampak buruk lockdown bagi masyarakat. sampai kapan begini pak??				
Sesudah				
kenaikan	kasus	sampai	covid	yuk
kematian	terus	positif	untuk	pak
meningkat	buruk	jumlah	bagi	cari
membantu	hari	sampai	kita	segala
masyarakat	hari	dampak	pak	kapan
lockdown	ke	solusi	begini	
melihat	yg	kapan	mari	

4.2.4. *Normalization*

Selanjutnya ialah normalization atau proses untuk menyesuaikan kata yang tidak sesuai berdasarkan KBBI seperti kata singkatan. Proses ini menggunakan operator dari rapidminert serta kamus. Kamus yang digunakan ialah kamus yang dibuat sendiri oleh peneliti berdasarkan temuan kata yang tidak sesuai pada dataset dan berdasarkan KBBI dimana kamus tersebut berformat txt [26]

Tabel 7 Contoh Normalization

Sebelum
yth kita sluruh rakyat indonesia dgn ini sy mengajak sluruh rakyat bersatu membuat indonesia bebas covid sdh cukup yg sakit yg phk dsb mari lakukan aturan pemerintah
Sesudah
yang terhormat kita seluruh rakyat indonesia dengan ini saya mengajak seluruh rakyat bersatu membuat indonesia bebas covid sudah cukup yang sakit yang pemutusan hubungan kerja dan sebagainya mari lakukan aturan pemerintah. mari cari solusi untuk membantu segala dampak buruk lockdown bagi masyarakat. sampai kapan begini pak??

4.2.5. *Stopword*

Pada proses stopword ini dataset akan dilakukan pemfilteran kata yang tidak memilki makna seperti kata sambung. Langkah ini menggunakan operator stopword yang terdapat di rapidminer dan kamus. Kamus yang digunakan ialah gabungan kamus stopword yang didapatkan dari hasil penelitian [25] serta kamus yang dibuat oleh peneliti berdasarkan temuan kata yang tidak bermakna dari dataset dan berformat ".txt"

Tabel 8 Contoh Stopword

Sebelum
yang terhormat kita seluruh rakyat indonesia dengan ini saya mengajak seluruh rakyat bersatu membuat indonesia bebas covid sudah cukup yang sakit yang pemutusan hubungan kerja dan sebagainya mari lakukan aturan pemerintah
Sesudah
terhormat rakyat indonesia mengajak rakyat bersatu indonesia bebas covid sakit pemutusan hubungan kerja mari lakukan aturan pemerintah

4.2.6. *Stemming*

Pada tahap ini pesan teks pada dataset dilakukan proses untuk mengubah kata token yang memiliki imbuhan menjadi kata bentuk dasarnya berdasarkan KBBI [26]. tahap ini menggunakan operator di rapidminer dan kamus. Kamus file yang digunakan berformat ".txt" yang dibuat oleh peneliti berdasarkan temuan kata yang memiliki imbuhan dan diubah menjadi kata dasarnya berdasarkan KBBI.

Tabel 9 Contoh Stemming

Sebelum
terhormat rakyat indonesia mengajak rakyat bersatu indonesia bebas covid sakit pemutusan hubungan kerja mari lakukan aturan pemerintah
Sesudah
hormat rakyat indonesia ajak rakyat satu indonesia bebas covid sakit putus hubungan kerja mari lakukan atur pemerintah

4.3. *Pembobotan Kata*

Setelah dataset diolah melalui tahap preprocessing, hasil tersebut akan dilakukan pembobotan kata atau perhitungan bobot tiap kata pada setiap pesan teks dengan metode TF-IDF menggunakan operator pada rapidminer.

Word	Attribute Name	Total Occurences ↓	Document Occurences
pemerintah	pemerintah	1022	891
covid	covid	1016	902
indonesia	indonesia	957	855
virus	virus	181	157
tangan	tangan	178	169
rakyat	rakyat	157	125

Gambar 3 Contoh Hasil TF-IDF

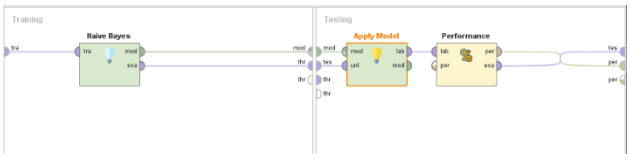
Dapat dilihat contoh gambar 3 dimana menghasilkan total frekuensi kemunculan pada dataset kata serta jumlah frekuensi dokumen atau pesan teks yang muncul dari kata tersebut

4.4. *Implementas Algoritma*

Pada proses ini hasil dari pembobotan kata sebelumnya akan dilakukan penerapan algoritma klasifikasi yaitu naïve bayes classifier dan KNN

4.4.1. *Implementasi Algoritma NBC*

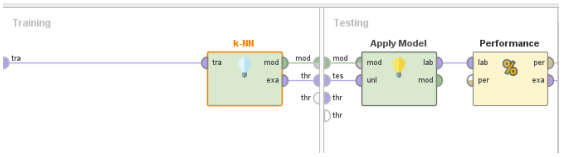
Pada proses implementasi algoritma naïve bayes ini menggunakan operator naïve bayes yang terdapat di rapidminer untuk klasifikasi di mana pada proses ini menggunakan dataset yang telah melalui proses preprocessing dan pembobotan kata



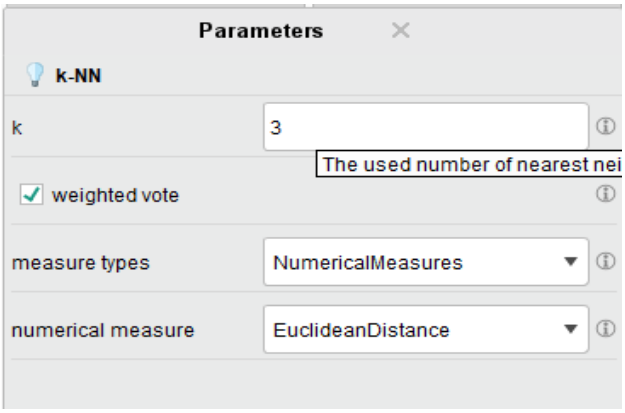
Gambar 4 Rangkaian Operator NBC

4.4.2. *Implementasi Algoritma KNN*

Pada tahap ini dilakukan implementasi algoritma KNN pada dataset hasil preprocessing dan pembobotan kata. Dimana implementasi menggunakan operator pemodelan KNN yang terdapat di rapidminer.



Gambar 5 Rangkaian Operator KNN

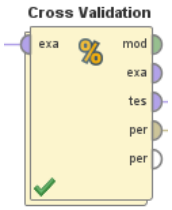


Gambar 6 Parameter Operator KNN

Pada pengaturan operator terdapat parameter yang akan diatur diantaranya ialah "k" yaitu jumlah dokumen terdekat dengan data uji yang akan diambil. Measure types yaitu tipe atribut yang akan dihitung dimana dipilih NumericalMeasures karena data yang dihasilkan pada proses pembobotan kata ialah numerik. Kemudian parameter numerical measure dipilih EucludianDistance karena rumus ini ialah perhitungan jarak yang paling sering dipakai pada dataset angka [19]. Menurut [27] nilai k sendiri ditentukan dengan angka ganjil dan lebih besar dari jumlah kelas, karena jika dipilih dengan angka genap maka akan ada kemungkinan hasil klasifikasi sulit untuk ditentukan karena akan terdapat hasil klasifikasi yang nilai masing-masing kelasnya sama. Maka dari itu nilai k pada penelitian ini ialah 3 karena telah cukup untuk memenuhi syarat yang sudah disebutkan sebelumnya.

4.5. Validasi

Pada langkah dilakukannya validasi menggunakan *K-Fold Validation* dimana dilakukan percobaan 10 kali dan membagi dataset menjadi 10 bagian secara acak dengan komposisi kelas yang sama alasan dilakukannya percobaan sebanyak 10 kali ialah karena semakin banyak percobaan dilakukan serta semakin banyak data yang dilatih semakin baik juga akurasi yang didapatkan [28]. Proses ini menggunakan operator *Cross Validation* yang terdapat di rapidminer, dimana dalam *Cross Validation* akan menampung rangkaian operator NBC dan KNN.



Gambar 7 Operator Cross Validation

4.6. Evaluasi

Pada tahap ini dilakukan evaluasi confusion matrix dari hasil pengujian kedua algoritma klasifikasi NBC dan KNN

4.6.1. Confusion Matrix NBC

accuracy: 67.84% +/- 4.04% (micro average: 67.85%)

	true pos	true neg	class precision
pred. pos	275	174	61.25%
pred. neg	117	339	74.34%
class recall	70.15%	66.08%	

Gambar 8 Confusion Matrix NBC

Langkah dilakukannya evaluasi *performance* penerapan metode NBC. Dimana pada confusion matrix akurasi yang didapatkan sebesar 67.84%, TP atau label positif yang diprediksi benar sebanyak 275 sedangkan FP atau label negatif yang diprediksi salah sebanyak 174 dan TN atau label negatif yang diprediksi benar sebanyak 339 sedangkan FN atau label positif yang diprediksi salah sebanyak 117.

4.6.2. Confusion Matrix KNN

accuracy: 72.37% +/- 3.72% (micro average: 72.38%)

	true pos	true neg	class precision
pred. pos	262	120	68.59%
pred. neg	130	393	75.14%
class recall	66.84%	76.61%	

Gambar 9 Confusion Matrix KNN

Tahap ini dilakukan evaluasi performa penerapan algoritma KNN. Dimana pada hasil confusion matrix akurasi yang didapatkan sebesar 72.37%, TP atau label positif yang diprediksi benar sebanyak 262 sedangkan FP atau label negatif yang diprediksi salah sebanyak 120 dan TN atau label negatif yang diprediksi benar sebanyak 393 sedangkan FN atau label positif yang diprediksi salah sebanyak 130.

4.7. Perbandingan Hasil

Pada tahap ini dilakukan perbandingan hasil confusion matrix antara hasil confusion matrix NBC dan hasil confusion matrix KNN dimana disajikan dalam bentuk sebagai berikut.

Tabel 10 Perbandingan Confusion Matrix 1

	Accuracy	Precision		Recall	
		pos	neg	pos	neg
NBC	67.84%	61.25%	74.34%	70.15%	66.08%
KNN	72.37%	68.59%	75.14%	66.84%	76.61%

Tabel 11 Perbandingan Confusion Matrix 2

	TP	TN	FP	FN
NBC	275	339	174	117
KNN	262	393	120	130

Dari perbandingan hasil confusion matrix pada tabel 10 dan 11 didapatkan bahwa perolehan akurasi pada penerapan algoritma KNN lebih besar 4.53% dibanding algoritma NBC. Dimana didapatkan akurasi tertinggi yaitu 72.37% dengan total sebanyak 905 data dimana 392 sentimen positif dan 513 sentimen negatif. Presisi sentimen postif KNN lebih besar 7.34% dan presisi sentimen negatif KNN lebih besar 0.80% dibanding NBC, dimana ini disebabkan karena lebih besarnya TN pada presisi negatif dan lebih kecilnya FP pada presisi positif dibanding NBC. Recall sentimen negatif KNN lebih besar 10.53% yang disebabkan oleh lebih besarnya TN dan lebih kecilnya FP KNN dibanding NBC. Sedangkan Recall sentimen positif KNN lebih kecil 3.31% dibanding NBC yang disebabkan oleh lebih besarnya TP dan lebih kecilnya FN NBC dibanding KNN

5. KESIMPULAN DAN SARAN

Hasil penelitian yang dilakukan pada dataset kumpulan opini pengguna twitter tentang covid-19 maka dapat disimpulkan dari hasil perbandingan penerapan algoritma klasifikasi KNN dan NBC pada analisis sentimen dari dataset kumpulan opini pengguna twitter tentang *COVID-19* pada tabel 4.14 bahwa algoritma KNN mendapat akurasi sebesar 72.37% sedangkan algoritma NBC mendapat akurasi sebesar 67.84%. KNN menjadi algoritma klasifikasi yang paling baik untuk mengklasifikasikan sentimen(pendapat) negatif, Dimana label negatif yang diprediksi benar(TN) pada KNN lebih besar yaitu 393 dibanding NBC yang sebanyak 339. sedangkan NBC menjadi algoritma yang paling baik untuk mengklasifikasikan sentimen(pendapat) positif, dimana label positif yang diprediksi benar(TP) pada NBC lebih besar yaitu 275 dibanding KNN yang sebanyak 262 berdasarkan hasil dari analisis sentimen(pendapat) pada dataset kumpulan opini pengguna twitter tentang covid-19. Saran untuk penelitian selanjutnya ialah : Penelitian dapat ditingkatkan lebih baik dengan menerapkan algoritma klasifikasi lain sebagai pembanding algoritma. Penelitian dapat ditingkatkan lebih baik dengan menerapkan teknik stemming lain seperti algoritma porter, algoritma nazief dan adriani dan lain-lain yang tidak hanya stemming secara manual

DAFTAR PUSTAKA

[1] S. Kemp, "DIGITAL 2021: INDONESIA," *DATA REPORTAL*, 2021. https://datareportal.com/reports/digital-2021-indonesia (accessed Dec. 27, 2021).

[2] M. I. Fikri, T. S. Sabrila, Y. Azhar, and U. M. Malang, "Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter," *SMATIKA JURNAL*, vol. 10, no. 02, pp. 71–76, 2020.

[3] M. A. Shereen, S. Khan, A. Kazmi, N. Bashir, and R. Siddique, "COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses," *Journal of Advanced Research*, vol. 24. Elsevier B.V., pp. 91–98, Jul. 01, 2020. doi: 10.1016/j.jare.2020.03.005.

[4] M. P. Lestari, D. J. Witarsyah, and F. Hamami, "PERAMALAN PERTAMBAHAN PASIEN COVID-19 MENGGUNAKAN SUPPORT VECTOR REGRESSION FORECASTING GROWTH OF COVID-19 PATIENTS USING SUPPORT VECTOR REGRESSION."

[5] E. Kartika Sari, B. Ria EMD, M. Karina Putri, M. Eka Rosita, P. S. Studi, and S. Tinggi Ilmu Kesehatan Akbidyo, "PANGAN FUNGSIONAL SEBAGAI ALTERNATIF PENUNJANG IMUN DI MASA PANDEMI," 2021.

[6] I. Mahfud and A. Gumantan, "Survey Of Student Anxiety Levels During The Covid-19 Pandemic," *Jp.jok (Jurnal Pendidikan Jasmani, Olahraga dan Kesehatan)*, vol. 4, no. 1, pp. 86–97, Nov. 2020, doi: 10.33503/jp.jok.v4i1.1103.

[7] E. T. Handayani and A. Sulistiyawati, "ANALISIS SENTIMEN RESPON MASYARAKAT TERHADAP KABAR HARIAN COVID-19 PADA TWITTER KEMENTERIAN KESEHATAN DENGAN METODE KLASIFIKASI NAIVE BAYES," *Jurnal Teknologi dan Sistem Informasi (JTSI)*, vol. 2, no. 3, pp. 32–37, 2021, [Online]. Available: http://jim.teknokrat.ac.id/index.php/JTSI

[8] R. Dwi Septiana, A. Budi Susanto, and T. Tukiyat, "Analisis Sentimen Vaksinasi Covid-19 Pada Twitter Menggunakan Naive Bayes Classifier Dengan Feature Selection Chi-Squared Statistic Dan Particle Swarm Optimization," pp. 49–56, 2021.

[9] N. M. A. J. Astari, Dewa Gede Hendra Divayana, and Gede Indrawan, "Analisis Sentimen Dokumen Twitter Mengenai Dampak Virus Corona Menggunakan Metode Naive Bayes Classifier," *Jurnal Sistem dan Informatika (JSI)*, vol. 15, no. 1, pp. 27–29, Nov. 2020, doi: 10.30864/jsi.v15i1.332.

[10] P. H. Prastyo, A. S. Sumi, A. W. Dian, and A. E. Permanasari, "Tweets Responding to the Indonesian Government's Handling of COVID-19: Sentiment Analysis Using SVM with Normalized Poly Kernel," *Journal of Information Systems Engineering and Business Intelligence*, vol. 6, no. 2, p. 112, Oct. 2020, doi: 10.20473/jisebi.6.2.112-122.

[11] M. A. Rofiqi, Abd. C. Fauzan, A. P. Agustin, and A. A. Saputra, "Implementasi Term-Frequency Inverse Document Frequency (TF-IDF) Untuk Mencari Relevansi Dokumen Berdasarkan Query," *ILKOMNIKA: Journal of Computer Science and Applied Informatics*, vol. 1, no. 2, pp. 58–64, Dec. 2019, doi: 10.28926/ilkomnika.v1i2.18.

[12] T. N. Wijaya, R. Indriati, and M. N. Muzaki, "Analisis Sentimen Opini Publik Tentang Undang-Undang Cipta Kerja Pada Twitter," *Jambura Journal of Electrical and Electronics Engineering*, vol. 2, no. 2, pp. 78–83, 2021.

[13] S. E. Saad and J. Yang, "Twitter Sentiment Analysis Based on Ordinal Regression," *IEEE Access*, vol. 7, pp. 163677–163685, 2019, doi: 10.1109/ACCESS.2019.2952127.

[14] R. Kurniawan and A. Apriliani, "Analisis sentimen masyarakat terhadap virus corona berdasarkan opini dari Twitter berbasis web scraper," *Jurnal Instek (Informatika Sains dan Teknologi)*, vol. 5, no. 1, pp. 67–75, 2020.

[15] E. Kartika and J. Gondohanindijo, "Rancang Bangun Model Sentimen Analisis Review Produk Pada Toko Online Menggunakan Naive Bayes," *Seminar Nasional Hasil Penelitian dan Pengabdian Kepada Masyarakat*, pp. 201–212, 2020.

[16] E. Ramadhanta, M. Razaq, D. W. Jacob, and F. Hamami, "Analisis Sentimen Kepuasan Mahasiswa Terhadap Pembelajaran Online Selama Pandemi Covid-19 Pada Media Sosial Twitter Menggunakan Perbandingan Algoritma Klasifikasi," *e-Proceeding of Engineering*, vol. 8, no. 5, pp. 9000–9006, 2021.

[17] J. A. Septian, T. M. Fahrudin, and A. Nugroho, "Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor," 2019. [Online]. Available: https://t.co/9WloaWpfD5

[18] Rafiqa Cahyani, "ANALISIS SENTIMEN PADA MEDIA SOSIAL TWITTER TERHADAP TOKOH PUBLIK PESERTA PILPRES 2019," SKRIPSI, UIN SUNAN AMPEL SURABAYA, Surabaya, 2019.

[19] N. Krisandi, B. Helmi, and I. Prihandono, "ALGORITMA k-NEAREST NEIGHBOR DALAM KLASIFIKASI DATA HASIL PRODUKSI KELAPA SAWIT PADA PT. MINAMAS KECAMATAN PARINDU," vol. 02, no. 1, p. 50, 2013.

[20] P. Pitria, "Analisis sentimen pengguna twitter pada akun resmi samsung indonesia dengan menggunakan naïve bayes ," *Doctoral dissertation*, 2014.

[21] A. Kurniawan and S. Adinugroho, "Analisis Sentimen Opini Film Menggunakan Metode Naïve Bayes dan Lexicon Based Features," 2019. [Online]. Available: http://j-ptiik.ub.ac.id

[22] Z. Muhammad, R. Ramadhani, and H. Rizqifaluthi, "Process Mining Akademik Sekolah Menggunakan RapidMiner," *MATICS*, vol. 10, no. 2, p. 39, Mar. 2018, doi: 10.18860/mat.v10i2.5745.

[23] Ph. D. Zainal A. Hasibuan, *METODOLOGI PENELITIAN PADA BIDANG ILMU KOMPUTER DAN TEKNOLOGI INFORMASI*. 2007.

[24] Prof. Dr. Sugiyono, *METODE PENELITIAN KUANTITATIF, KUALITATIF DAN R & D*. ALFABETA, 2013.

[25] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," 2003.

[26] Kemdikbud, "KBBI Daring," Oct. 28, 2016. https://kbbi.kemdikbud.go.id/ (accessed Dec. 11, 2022).

[27] M. Rivki and A. M. Bachtiar, "IMPLEMENTASI ALGORITMA K-NEAREST NEIGHBOR DALAM PENGKLASIFIKASIAN FOLLOWER TWITTER YANG MENGGUNAKAN BAHASA INDONESIA," *Jurnal Sistem Informasi*, vol. 13, no. 1, p. 31, May 2017, doi: 10.21609/jsi.v13i1.500.

[28] A. A. Nabhan, B. Rahayudi, and D. E. Ratnawati, "Klasifikasi Tweets Masyarakat yang Membicarakan Layanan GoFood dan GoRide pada GoJek Dimedia Sosial Twitter Selama Masa Kenormalan Baru (New Normal) dengan Metode Naïve Bayes," 2021. [Online]. Available: http://j-ptiik.ub.ac.id

BIODATA PENULIS

Habibi Aulia Nur Syifa
Merupakan Mahasiswa Jurusan Sistem Informasi di Universitas Nusantara PGRI Kediri.

Arie Nugroho
Merupakan Dosen Di Universitas Nusantara PGRI Kediri Program Studi Sistem Informasi.

Rina Firliana
Merupakan Dosen Di Universitas Nusantara PGRI Kediri Program Studi Sistem Informasi