JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1545-1553
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4220

Perbandingan Metode Naïve Bayes dan Support Vector Machine Untuk Analisis Sentimen Terhadap Vaksin Astrazeneca di Twitter

Eva Rahma Indriyani, Paradise, Merlinda Wibowo*

Fakultas Informatika,Teknik Informatika, Institut Teknologi Telkom Purwokerto, Banyumas, Indonesia
Email: [1]18102011@ittelkom-pwt.ac.id, [2]paradise@ittelkom-pwt.ac.id, [3,*]merlinda@ittelkom-pwt.ac.id
Email Penulis Korespondensi: merlinda@ittelkom-pwt.ac.id

Abstrak−Pelaksanaan vaksinasi Covid-19 di Indonesia mendapatkan berbagai opini yang pro dan kontra dari masyarakat. Ditemukanya disinformasi dan misinformasi tentang vaksin melalui konten media sosial dapat mempengaruhi serapan informasi seseorang sehingga mengarah pada penundaan vaksin. Padahal vaksinasi adalah kontribusi terbesar dan paling efektif untuk mengakhiri pandemi covid-19. Salah satu vaksin yang disediakan oleh pemerintah Indonesia adalah Astrazeneca. Vaksin Astrazeneca sempat menjadi perdebatan di masyarakat karena kehalalan dan kemanannya karena isu ditemukanya unsur tripsin babi dalam vaksin tersebut. Twitter saat ini telah menjadi wadah bagi para pengguna untuk mengungkapkan kekhawatiran dan opini terhadap vaksin Covid-19. Pada penelitian ini, digunakan *llibrary* snscrape untuk mengumpulkan data dari periode Mei sampai Juni 2021. Dataset berjumlah sebanyak 3105 tweet. Dataset yang telah dikumpulkan kemudian dilakukan preprocessing untuk mengoptimalkan data. Setelah melewati tahap preprocessing, data dilakukan pelabelan kelas tweet menggunakan kamus *lexicon based* yang menghasilkan 1275 tweet dengan label opini positif dan 1830 tweet berlabel opini negatif. Penelitian ini mengkaji kinerja Naïve Bayes dan Support Vector Machine dengan menambahkan teknik TF-IDF (*Term Frequency - Inverse Document Frequency*) yang bertujuan untuk menambah bobot suatu kata pada dokumen. Hasil evaluasi menunjukkan Support Vector Machine memiliki akurasi, presisi, recall dan f1-score yang lebih besar yaitu 87.27%, 90.41%, 77,34% dan 83.37% dibandingkan Naïve Bayes yang memiliki akurasi, presisi, recall dan f1-score sebesar 76.81%, 72.40%, 70.70% dan 71.52%.

Kata Kunci: Analisis Sentimen; Vaksin Astrazeneca; Support Vector Machine; Naïve Bayes; Twitter

Abstract−The implementation of Covid-19 vaccination in Indonesia turned out to have various pro and contra opinions from the public. The discovery of disinformation and misinformation about vaccines spread through social media content affects a person's absorption of information so which leads to vaccine delays. When in fact, vaccination is one of the biggest and most effective contributions to preventing the Covid-19 pandemic. Astrazeneca is one of the vaccines provided by the Indonesian government. This vaccine used to be controversial amongst the public regarding its halalness and the safety of the vaccine because of the issue of the said vaccine containing swine trypsin. Nowadays Twitter has become a place for users to express their concerns and opinion regarding the Covid-19 vaccine. Data obtained from Twitter will be useful if it is analyzed, one of which is sentiment analysis. In this study, data collection was carried out using the snscrape library with a total of 3105 tweets obtained from the period May to June 31, 2021. The dataset that has been collected is then preprocessed to optimize the data. After passing the preprocessing stage, the data was labeled as tweet class using a lexicon-based dictionary which resulted in 1275 tweets with positive opinin labels and 1830 tweets labeled as negative opinion. The aim of this study is to examines the performance of Naïve Bayes and Support Vector Machine with adding the weighting method TF-IDF (*Term Frequency – Inverse Document Frequency*). The evaluation results show that the Support Vector Machine has a greater accuracy, precision, recall and f1-score of 87.27%, 90.41%, 77,34% and 83.37% compared to Naïve Bayes which has an accuracy, precision, recall and f1- of 76.81%, 72.40%, 70.70% and 71.52%.

Keywords: Sentiment Analysis; Astrazeneca Vaccine; Support Vector Machine; Naïve Bayes; Twitter.

1. PENDAHULUAN

Pada tanggal 11 Maret 2020, World Health Organization (WHO) meresmikan sebuah klaster pneumonia yang dikenal sebagai COVID (Coronavirus Disease) sebagai pandemi global dengan jumlah kasus sebanyak lebih dari 243 juta jiwa yang telah terinfeksi dan jumlah kematian sebanyak 4,9 juta jiwa per tanggal 24 Oktober 2021. Pemerintah Indonesia telah memberlakukan beberapa kebijakan dan peraturan mengenai protokol kesehatan serta pembatasan sosial untuk menekan penyebaran Covid-19. Namun, meningkatkan kekebalan dengan vaksinasi merupakan kontribusi terbesar dan paling efektif dalam mengakhiri pandemi. Dibutuhkan setidaknya 70% dari total populasi yang harus di vaksinasi untuk mencapai sebuah herd immunity [1]. Vaksin AstraZeneca adalah vaksin covid-19 yang diproduksi di Universitas Oxford dengan menggunakan replika adenovirus simpanse yang tidak sempurna sebagai vector. Vaksin AstraZeneca sempat menjadi perdebatan tentang kehalalan dan kemanannya karena isu ditemukanya unsur babi dalam vaksin tersebut. Namun vaksin yang diciptakan oleh University of Oxford justru menjadi rekomendasi WHO dengan efikasi sebesar 76%, lebih besar dibandingkan vaksin Sinovac yang hanya mencapai 63.50% [2]. Pelaksanaan vaksinasi di Indonesia mendapatkan berbagai macam tanggapan dan opini dari masyarakat [3]. Berdasarkan survei secara *online* yang dilakukan oleh kementrian Kesehatan Republik Indonesia, ITAGI (Indonesian Technical Advisory Group on Immunization), UNICEF (United Nations International Children's Emergency Fund) dan WHO terhadap 76 % responden berusia 18-45 pada 19-30 September 2020, diketahui ada kekhawatiran dan ketidakpercayaan terhadap efektivitas serta kehalalan vaksin [4].

Konten terkait dengan vaksin telah banyak tersebar di media sosial [5]. Menurut data yang diperoleh dari Global Digital Statistic “Digital, Social & Mobile in 2021” di We Are Social (2021), total pengguna media sosial di Indonesia berjumlah lebih dari 170 juta pengguna. Pengguna twitter sendiri mencakup lebih dari 63% dari

Submitted: **08/06/2022**; Accepted: **05/07/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1545-1553
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4220

total pengguna media sosial di Indonesia serta menduduki peringkat-5 media sosial yang paling sering digunakan di tahun 2021. Data dari twitter penting dan bermanfaat bagi masyarakat atau organisasi jika dianalisa. Salah satunya adalah dengan analisis sentimen. Dengan analisis sentimen, polaritas dari sebuah opini dapat digunakan untuk memprediksi suasana publik.

Analisis Sentimen adalah salah satu cara untuk memahami suatu data tekstual secara otomatis dengan cara mengekstraksi, mengolah data [6]. Ada dua metode yang dapat digunakan untuk menganalisis sentimen dari suatu dokumen diantaranya yaitu dengan menggunakan pendekatan Lexicon dan Machine Learning. Klasifikasi menggunakan pendekatan lexicon merupakan metode klasifikasi yang berdasarkan data positif ataupun kata negatif yang ada pada data yang kemudian dicocokkan dengan kamus lexicon. Untuk metode Machine Learning, adalah sebuah metode dengan memanfaatkan pembelajaran dari data latih. Algoritma Naïve Bayes serta Support Vector Machine merupakan algoritma klasifikasi machine learning yang umumnya paling banyak digunakan oleh para peneliti di bidang *text mining* [7].

Penelitian terdahulu tentang Analisis Sentimen adalah tentang Opini pengguna Gopay [8]. Studi ini menggunakan metode SVM untuk mengklasifikasikan sentimen serta membandingkan kernel linear dan kernel polynomial. Dari penelitian ini menghasilkan bahwa metode Support Vector Machine dengan kernel linear memiliki akurasi yang lebih tinggi yaitu sebesar 89.17% daripada menggunakan kernel polynomial yang hanya sebesar 84.38%.

Penelitian tentang Analisis selanjutnya menggunakan data dari opini public terhadap Undang Undang Cipta Kerjna [9]. Peneliti menggunakan algoritma Naive Bayes Classifier dengan pembobotan TF-IDF. Data yang diperoleh berjumlah 1000 tweet dengan periode tweet yang dikirim dari Oktober 2020 sampai November 2020. Pelabelan dilakukan secara manual untuk mengkategorikan sentimen positif dan negatif. Dari hasil pengujian yang telah dilakukan, diperoleh bahwa performa algoritma Naive Bayes dengan pembagian rasio data training 80% dan data testing 20% memiliki nilai akurasi terbaik yaitu 89,9%.

Dalam penelitian yang lain di tahun 2021 yang berjudul Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19, dilakukan perbandingan antara metode Naïve Bayes dan Support Vector Machine [10]. Data yang digunakan dalam penelitian ini berjumlah sebanyak 845 tweet yang diambil dari twitter menggunakan teknik webscraping dengan tools Octoparse berdasarkan dua kata kunci, yaitu ‘vaksinmerahputih’ dan ‘vaksinsinovac’. Hasil dari penelitian ini menunjukkan bahwa penguunaan algoritma Naïve Bayes dengan evaluasi 10 k-fold cross validation mempunyai nilai akurasi terbaik dibandingkan metode Support Vector Machine dengan persentase akurasi sebesar 85,59%.

Penelitian selanjutnya yaitu tentang Analisis Sentimen Pada Tweet Terkait Vaksin Covid-19 Menggunakan Metode Support Vector Machine. Pada penelitian ini dilakukan klasifikasi dengan menggunakan metode Support Vector Machine serta membandingkan tokenisasi unigram dan bigram. Hasil dari penelitian ini, diperoleh bahwa perbandingan ukuran evaluasi tokenisasi menggunakan unigram dan bigram tidaklah jauh. Nilai tertinggi akurasi seluruh aspek pengukuran yaitu accuracy, precision, recall dan f1-score adalah 84%

Telah dilakukan juga penelitian di tahun 2021 juga yang membahas tentang analisis sentimen mengenai vaksin yang berjenis Sinovac dengan data yang berasal dari media sosial twitter [11]. Dalam penelitian ini, data diambil menggunakan metode crawling dengan python. Proses klasifikasi data dalam penelitian ini menggunakan metode Naïve Bayes yang dikategorikan menjadi 2 jenis polaritas sentimen yaitu positif dan negatif. Hasil dari penelitian ini menunjukkan bahwa Naïve Bayes mampu mengklasifikasikan tweet dan mengolompokkan nya dalam sentimen positif dan negatif. Probabilitas dari sentimen positif memiliki hasil yang lebih besar daripada sentimen negatif. Sehingga sebuah respon dengan komentar positif memiliki probabilitas lebih besar kemungkinan respon dengan komentar negatif.

Penelitian ini akan dilakukan perbandingan menggunakan dua metode yaitu Naïve Bayes dan Support vector Machine. Berdasarkan penelitian yang telah dilakukan sebelumnya,. Metode Naïve Bayes merupakan salah satu algoritma machine learning yang biasa diterapkan dalam analisis sentimen. Disisi lain, Support Vector Machine juga merupakan salah satu algoritma yang tepat digunakan untuk text classification. Kemampuan Support Vector Machine menemukan hyperplane yang optimal menjadikan algoritma ini memiliki tingkat generalitas yang tinggi, menjadikannya algoritma yang paling akurat dibandingkan algoritma lainnya [12]. Meskipun penelitian ini menggunakan metode yang sudah ada, namuan kajian penelitian tetap dibutuhkan terutama konfigurasi yang untuk membuat model klasifikasi untuk analisis sentimen.

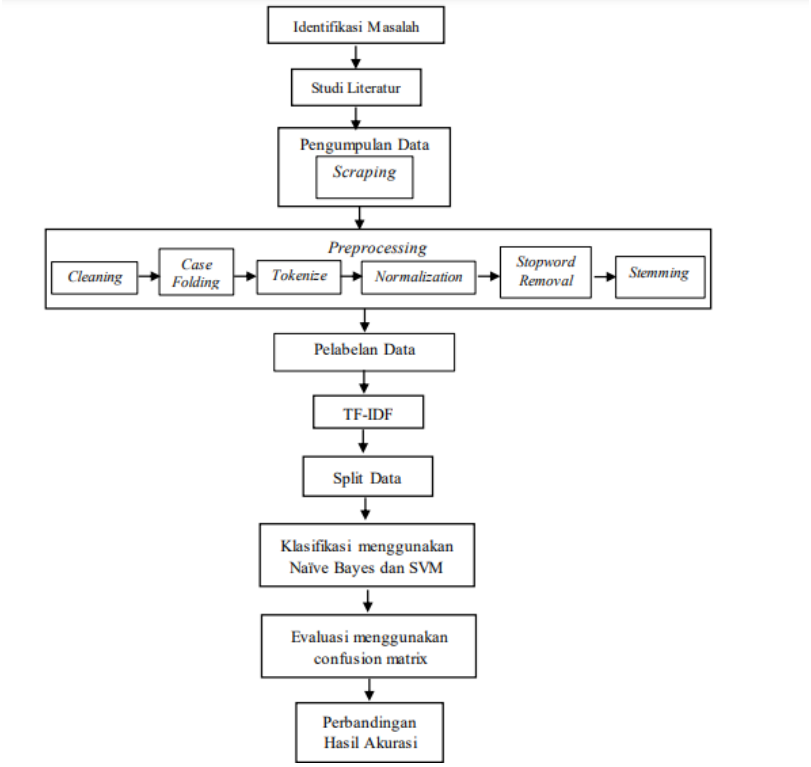
2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Tahapan penelitian yang dilakukan pada penelitian ini adalah scraping data, preprocessing data, ekstraksi fitur, pelabelan data, pembangunan model klasifikasi, dan kemudian diakhiri dengan evaluasi performa klasifikasi. Pada Gambar 1 berikut merupakan alur penelitian akan dilakukan:

Submitted: **08/06/2022**; Accepted: **05/07/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1545-1553
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4220



Gambar 1. Tahapan Penelitian

2.2 Pengumpulan Data

Data yang dikumpulkan berupa data teks yang diambil dengan teknik scraping. Pengumpulan data yang digunakan pada penelitian ini adalah opini atau tweet pada Twitter yang berkaitan dengan topik vaksin Astrazeneca. Data yang dikumpulkan berdasarkan pencarian kata kunci 'vaksin Astrazeneca' pada periode bulan Mei sampai Juni 2021 di Twitter.

2.3 Preprocessing Data

Preprocessing adalah proses perubahan dari yang awalnya data tidak terstruktur menjadi data yang terstruktur. Preprocessing merupakan langkah terpenting dalam data mining [13]. Proses pembersihan yang dilakukan pada tahap ada 6 tahap, diantaranya adalah: *cleaning*, *case folding*, *tokenization*, *normalization*, *stopword removal*, dan *stemming*.

2.4 Pelabelan Data

Langkah selanjutnya adalah pelabelan data. Tahap *Labelling* dalam penelitian ini menggunakan pendekatan lexicon. Kamus Lexicon yang akan dipakai adalah kamus Lexicon Inset dari penelitian yang dilakukan berdasarkan penelitian Fajri (2017) [14] yang berjudul "Evaluasi daftar kata untuk sentimen analisis berbahasa Indonesia". InSet Lexicon terdiri dari dua kamus lexicon yaitu lexicon negatif yang berisi 6.609 kata negatif dan lexicon positif yang berisi 3.609 kata positif. Setiap kata memiliki nilai bobot atau *polarity score* dengan rentang bobot antara -5 hingga +5. Nilai *Polarity score* ini yang nantinya akan digunakan untuk mengklasifikasikan jenis pelabelan untuk setiap data, masuk kedalam kelas label negatif atau positif [15]

2.4 Ekstraksi Fitur dengan TF-IDF

Metode TF-IDF adalah metode yang paling umum digunakan untuk menghitung bobot setiap kata dalam pencarian informasi. TF-IDF adalah ukuran statistik yang digunakan untuk menilai seberapa penting sebuah kata di dalam dokumen atau dalam sekelompok kata Konsep dari Term Frequency (TF) menunjukkan bahwa semakin sering suatu kata yang muncul pada sebuah dokumen maka semakin tinggi juga nilai bobot kata tersebut. Proses IDF *(Inverse Document Frequency)*, di sisi lain merupakan kebalikan dari TF karena semakin sering suatu term muncul dalam proses IDF maka semakin kecil nlai bobot dari term itu sendiri [16]. Untuk menghitung nilai IDF, digunakan persamaan 1:

$$idf_t = log \frac{N}{df_t} \quad (1)$$

Sementara untuk menghitung nilai TF-IDF, dilakukan dengan persamaan 2.2:

$$Tf.idf_{t,d} = tf_{t,d} \, x \, Idf_t \quad (2)$$

Submitted: **08/06/2022**; Accepted: **05/07/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1545-1553
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4220

Dimana:
idf_t = invers frekuensi dokumen dari kata t
N = banyaknya dokumen
df_t = banyaknya dokumen yang didalamnya terdapat kata t
$Tf.idf_{t,d}$ = nilai bobot kata t pada dokumen d
$tf_{t,d}$ = frekuensi kemunculan term t pada dokumen d

2.5 Klasifikasi Menggunakan Naïve Bayes dan SVM

Klasifikasi adalah salah satu dari teknik dari data mining yang termasuk dalam pembelajaran terbimbing (*supervised learning*). Tujuannya adalah untuk memprediksi target dari beberapa atribut.

1. Naïve Bayes merupakan metode klasifikasi berdasarkan teorema Bayes. Algoritma ini pertama kali dikemukakan oleh Thomas Bayes. Disebut naïve karena kondisi antar atribut diasumsikan saling bebas. Metode klasifikasi ini tepat digunakan saat masukan dalam jumlah yang besar [17].
   Rumus untuk Naive bayes classifier ada pada persamaan berikut:

$$P(c_i) = \frac{fd(c_i)}{|D|} \quad (3)$$

   Keterangan:
   $P(c_i)$ = Probabilitas c_i yang merupakan kategori kelas
   $fd(c_i)$ = Jumlah dokumen c_i
   $|D|$ = Jumlah data latih / dokumen

   Setelah mendapatkan probabilitas dari setiap kelas, selanjutnya yaitu menghitung probabilitas dari setiap fitur pada setiap kelas dengan persamaan:

$$P(w_k|c_i) = \frac{f(w_{ki},C)+1}{P(C_i)+|W|} \quad (4)$$

   Keterangan:
   $P(w_k|c_i)$ = Peluang kemunculan kata ada sebuah kelas, w_k adalah kata yang muncul pada sebuah kategori
   $f(w_{ki}, C)$ = Nilai kemunculan kata w_{ki} pada kelas c_i
   $P(C_i)$ = Jumlah keseluruhan kemunculan kata pada kelas c_i
   $|W|$ = Jumlah data latih/dokumen

   Kemudian langkah selanjutnya adalah menentukan probabilitas data uji dari setiap kelas berdasarkan dari proses pembelajaran. Nilai probabilitas yang paling tinggi akan terpilih:

$$Vmap = \underset{\{kelas\ 0, kelas\ 1\}}{argmax} \prod_{i=1}^{n} P(w_k|c_i)\ x\ \ P(c_i) \quad (5)$$

   Keterangan:
   $P(w_k|c_i)$ = Probabilitas kemunculan kata-kata pada sebuah kelas, w_k merupakan kata yang muncul pada sebuah kategori
   $P(c_i)$ = Menentukan probabilitas c_i yaitu kategori kelas
2. Support Vector Machine (SVM) adalah salah satu metode pembelajaran terbimbing yang menganalisis data dan mengenali pola yang digunakan untuk analisis klasifikasi.. SVM adalah algoritma yang mengimplementasikan fungsi hyperplane pada data untuk membentuk wilayah setiap kelas. Hyperplane merupakan sebuah fungsi yang digunakan sebagai pembatas antar kelas yang ada. Perhitungan batas hyperplane digunakan untuk mencari titik maksimum untuk mendapatkan garis hyperplane yang paling optimal saat memisahkan data menjadi dua kelas. Kernel yang akan digunakan dalam penelitian ini adalah fungsi kernel linear. Kernel linear dipilih karena merupakan kernel terbaik untuk data yang berup teks [18]. Fungsi linear kernel dinyatakan dengan persamaan berikut ini:

$$K(x_i, x_j) = (x_i, x_j) \quad (6)$$

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Data yang dikumpulkan berupa data teks yang diambil dengan teknik scraping menggunakan library Snscrape. Dipilihnya Snscrape untuk teknik scraping karena library ini tidak memiliki limit jumlah tweet ataupun rentang waktu ketika melakukan pengambilan data [19]. Data yang terkumpul sebanyak 6000 tweet dikumpulkan dan disimpan dalam format .json. Data yang didapatkan ini masih tercampur dengan bahasa lainya dan banyak data yang duplikat, maka dari itu telah dilakukan juga pembersihan data dengan menghapus tweet yang menggunakan bahasa asing atau tweet selain bahasa Indonesia, menghapus data yang kosong, serta menghapus data yang duplikat sehingga diperoleh data berjumlah 3105 tweet.

Submitted: **08/06/2022**; Accepted: **05/07/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1545-1553
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4220

Tabel 2. Contoh Data Hasil Scraping

No	Tweet
1	Alhamdulillah . Nikmat bener. Ikhtiar, biar ikutan andil untuk negara supaya lepas dari c19. Btw aku dapet vaksin AstraZeneca, kebetulan baru bisa kemarin.
2	mau vaksin ke 2 jadwal tgl 29,astrazeneca sampai sekarang ga ada kabarnya udah ada belum tuh vaksin nya?! serius ga sih ngurusin nya#vaksinastrazeneca#covid-19#DEPKESDKI#RSRESTI#.
3	@unmagnetism Aku udah vaksin astrazeneca dan ga ada efek apa2 ðŸ'ªðŸ'ªðŸ'ª Ayo vaksin gesss!!

3.2 Preprocessing Data

Preprocessing data dilakukan untuk menyeragamkan bentuk dan format agar dipersiapkan menjadi data yang dapat diolah pada tahap selanjutnya

a. Cleaning, merupakan proses pembersihan teks yang dilakukan dengan menghilangkan noise. Pada tahap ini, simbol, angka, dan tanda baca yang tidak diperlukan untuk analisis sentimen akan dihilangkan [20].

Tabel 3. Contoh Data Hasil Cleaning

Sebelum cleaning	Sesudah cleaning
mau vaksin ke 2 jadwal tgl 29,astrazeneca sampai sekarang ga ada kabarnya udah ada belum tuh vaksin nya?! serius ga sih ngurusin nya#vaksinastrazeneca#covid-19	mau vaksin ke jadwal tgl astrazeneca sampai sekarang ga ada kabarnya udah ada belum tuh vaksin nya serius ga sih ngurusin nya vaksinastrazeneca covid

b. Case Folding, yaitu merupakan tahap pengubahan huruf kapital pada data tweets menjadi huruf kecil list [21].

Tabel 4. Contoh Data Case Folding

Sebelum case folding	Sesudah case folding
mau vaksin ke jadwal tgl astrazeneca sampai sekarang ga ada kabarnya udah ada belum tuh vaksin nya serius ga sih	mau vaksin ke jadwal tgl astrazeneca sampai sekarang ga ada kabarnya udah ada belum tuh vaksin nya serius ga sih ngurusin

c. Tokenization, atau tokenisasi adalah proses memecah kalimat menjadi token dari teks tweet atau komponen komponen individual. Dengan dilakukan pemotongan akan memudahkan langkah selanjutnya karena pada proses-proses tersebut akan mencocokkan berdasarkan kata [22]

Tabel 5. Contoh Data Tokenisasi

Sebelum tokenisasi	Sesudah tokenisasi
mau vaksin ke jadwal tgl astrazeneca sampai sekarang ga ada kabarnya udah ada belum tuh vaksin nya serius ga sih ngurusin nya vaksinastrazeneca covid	['mau', 'vaksin', 'ke', 'jadwal', 'tgl', 'astrazeneca', 'sampai', 'sekarang', 'ga', 'ada', 'kabarnya', 'udah', 'ada', 'belum', 'tuh', 'vaksin', 'nya', 'serius', 'ga', 'sih', 'ngurusin', 'nya', 'vaksinastrazeneca', 'covid']

d. Normalisasi, berguna untuk mengubah kata yang tidak baku menjadi kata baku serta. Dalam proses ini, digunakan sebuah data yang berisi sebuah padanan kata tidak baku dan baku yang tersimpan dalam sebuah kamus [23]

Tabel 6. Contoh Data Normalisasi

Sebelum normalisasi	Sesudah normalisasi
['mau', 'vaksin', 'ke', 'jadwal', 'tgl', 'astrazeneca', 'sampai', 'sekarang', 'ga', 'ada', 'kabarnya', 'udah', 'ada', 'belum', 'tuh', 'vaksin', 'nya', 'serius', 'ga', 'sih', 'ngurusin', 'nya', 'vaksinastrazeneca', 'covid']	['mau', 'vaksin', 'ke', 'jadwal', 'tanggal', 'astrazeneca', 'sampai', 'sekarang', 'tidak', 'ada', 'kabarnya', 'sudah', 'ada', 'belum', 'itu', 'vaksin', 'nya', 'serius', 'tidak', 'sih', 'mengurus', 'nya', 'vaksin astrazeneca', 'covid']

e. Stopword adalah tahap penghapusan kata yang tidak relevan terhadap penentuan klasifikasi sentimen. Jika termasuk di dalam daftar kata stopword pada NLTK (*natural language tool kit*) maka kata-kata tersebut akan. Kata-kata yang tersisa di dalam dokumen dianggap sebagai kata-kata yang penting [23]

Submitted: **08/06/2022**; Accepted: **05/07/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1545-1553
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4220

Tabel 7. Contoh Data Stopword Removal

Sebelum normalisasi	Sesudah normalisasi
['mau', 'vaksin', 'ke', 'jadwal', 'tanggal', 'astrazeneca', 'sampai', 'sekarang', 'tidak', 'ada', 'kabarnya', 'sudah', 'ada', 'belum', 'itu', 'vaksin', 'nya', 'serius', 'tidak', 'sih', 'mengurus', 'nya', 'vaksinastrazeneca', 'covid']	['vaksin', 'jadwal', 'tanggal', 'astrazeneca', 'kabarnya', 'vaksin', 'serius', 'mengurus', 'vaksinastrazeneca', 'covid']
['aku', 'sudah', 'vaksin', 'astrazeneca', 'dan', 'tidak', 'ada', 'efek', 'apa', 'ayo', 'vaksin', 'gesss']	['vaksin', 'astrazeneca', 'efek', 'ayo', 'vaksin', 'gesss']

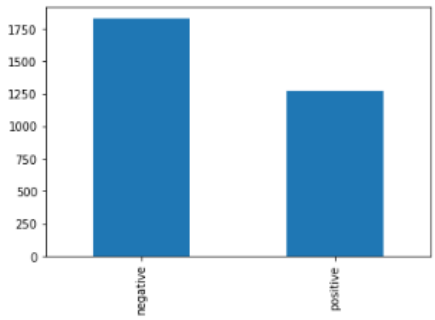
f. Stemming adalah tahap perubahan sebuah kata ke dalam bentuk kata dasarnya dengan cara menghilangkan imbuhan pada kata tersebut [24]

Tabel 8. Contoh Data Stemming

Sebelum normalisasi	Sesudah normalisasi
['vaksin', 'jadwal', 'tanggal', 'astrazeneca', 'kabarnya', 'vaksin', 'serius', 'mengurus', 'vaksinastrazeneca', 'covid']	['vaksin', 'jadwal', 'tanggal', 'astrazeneca', 'kabar', 'vaksin', 'serius', 'urus', 'vaksinastrazeneca', 'covid']
['vaksin', 'astrazeneca', 'efek', 'ayo', 'vaksin', 'gesss']	['vaksin', 'astrazeneca', 'efek', 'ayo', 'vaksin', 'gesss']

3.3 Pelabelan

Proses pelabelan pada penelitian ini yang dilakukan menggunakan pendekatan Lexicon. Pendekatan *Lexicon-based* bekerja dengan menggunakan kamus lexicon yang dilengkapi dengan bobot pada setiap katanya sebagai sumber leksikal. Penentuan kelas positif atau negatif didasari oleh nilai polarity score. Jika polarity score lebih dari 0 maka menunjukkan kelas sentimen positif, sedangkan jika polarity score mengarah ke nilai kurang dari 0 menunjukkan kelas sentimen negatif [15].



Gambar 2. Perbandingan data kelas negatif dan positif

Pada Gambar 2 menunjukkan bahwa pelabelan otomatis ini mendapatkan data sebanyak 1275 tweet bernilai positif dan 1830 tweet bernilai negatif.

3.4 Ekstraksi Fitur

Setelah data telah melewati tahap *labelling* dan *preprocessing*, maka tahap selanjutnya yaitu pembuatan fitur dengan TF-IDF TF-IDF adalah metode pembobotan dari hubungan suatu kata *term* terhadap istilah di setiap kalimat yang dianggap sebagai dokumen [25]. Implementasi TF-IDF menggunakan library python yang bernama sklearn.feature_extraction.text. Dari tahap TF-IDF diperoleh term unik sebanyak sebanyak 3643 kata.

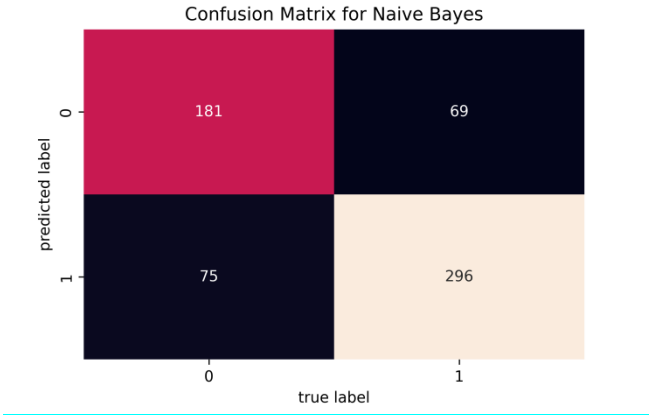
3.5 Analisis Metode Klasifikasi

Klasifikasi pada penelitian ini menggunakan dua metode yaitu Naïve Bayes Classifier (NBC) dan Support Vector Machine (SVM). Pada pengujian ini menggunakan pembagian data training dan data testing 80%:20% [26] dengan total data sebanyak 3105 yaitu 2484 data latih dan 621 data uji. Pada tahap evaluasi, digunakan metode confusion matrix. Dari confusion matrix akan diketahui nilai performa dari metode klasifikasi yang terdiri dari akurasi presisi, recall serta precision [27]

a. Analisis Klasifikasi Naïve Bayes
Hasil evaluasi model klasifikasi Naïve Bayes dengan confusion matrix dapat dilihat pada Gambar 3:

Submitted: **08/06/2022**; Accepted: **05/07/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1545-1553
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4220



Gambar 3. Visualisasi Confusion Matrix Naïve Bayes

Pada Gambar 4.14 memperlihatkan Confusion matrix berukuran 2 x 2 yang mewakili setiap kelas klasifikasi positif dan negatif . Dari confusion matrix dapat dijelaskan bahwa model mengklasifikasikan dengan benar sebesar 181 data positif, 296 data negatif. Sehingga, dari tabel confusion matrix diatas, maka diperoleh hasil klasifikasi dari Naïve Bayes dengan perhitungan sebagai berikut :

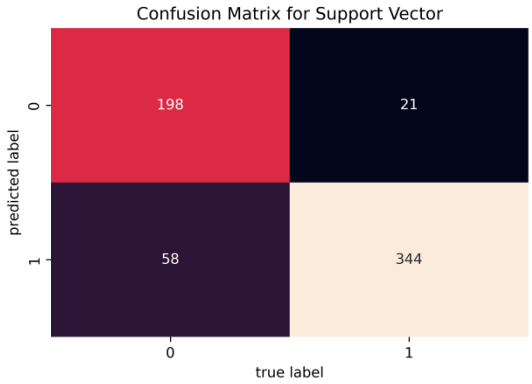
$$\text{Accuracy} = \frac{True\ Positive + True\ Negatif}{Total\ Data\ yang\ diuji}\ x\ 100\% = \frac{477}{621}\ x\ 100 = 76.81\%$$

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}\ x\ 100\% = \frac{181}{250}\ x\ 100\% = 72.4\%$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}\ x\ 100\% = \frac{181}{256}\ x\ 100\% = 70.7\%$$

$$\text{F1 -Score} = 2\ x\ \frac{Recall * Precision}{Recall + Precision}\ x\ 100\% = 2\ x\ \frac{70.7\% * 72.4\%}{70.7\% + 72.4\%}\ x\ 100\% = 71.52\%$$

b. Evaluasi Klasifikasi Support Vector Machine
Hasil evaluasi model klasifikasi Support Vector Machine dengan confusion matrix dapat dilihat pada Gambar 4:



Gambar 4. Visualisasi Confusion Matrix SVM

Pada gambar 4. memperlihatkan Confusion matrix berukuran 2 x 2 yang mewakili setiap kelas klasifikasi positif dan negatif. Dari confusion matrix dapat dijelaskan bahwa model mengklasifikasikan dengan benar sebesar 198 data positif dan 344 data negatif. Sehingga, dari tabel confusion matrix diatas, maka diperoleh hasil klasifikasi dari Support Vector Machine dengan perhitungan sebagai berikut :

$$\text{Accuracy} = \frac{True\ Positive + True\ Negatif}{Total\ Data\ yang\ diuji}\ x\ 100\% = \frac{542}{621}\ x\ 100\% = 87.27\%$$

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}\ x\ 100\% = \frac{198}{219}\ x\ 100\% = 90.41\%$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}\ x\ 100\% = \frac{198}{256}\ x\ 100\% = 77.34\%$$

$$\text{F1 Score} = 2\ x\ \frac{Recall + Precision}{Recall + Precision}\ x\ 100\% = 2\ x\ \frac{77.34\% * 90.4\%}{77.34\% + 90.4\%}\ x\ 100\% = 83.37\%$$

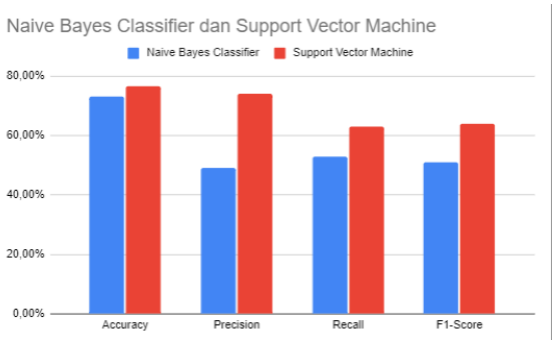
3.6 Perbandingan Metode Klasifikasi

Berdasarkan hasil analisis vaksin sentimen astrazeneca dengan algoritma Naïve Bayes dan Support Vector Machine melalui hasil preprocessing dan pembobotan term dengan TF-IDF serta evaluasi hasil klasifikasi berlabel positif dan negatif diperoleh bahwa akurasi dari metode Naïve Bayes adalah sebesar 76.81% dan SVM sebesar

Submitted: **08/06/2022**; Accepted: **05/07/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1545-1553
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4220

87.27%. Selanjutnya dapat dilihat nilai precision, recall dan f1-score dari setiap kelas klasifikasi. Untuk metode Naïve Bayes diperoleh *precision*, *recall*, dan *f1-score* sebesar 90.41%, 77,34% dan 83.37%. Sedangkan untuk metode Support Vector Machine memperoleh *precision*, *recall*, dan f1-score sebesar 72.40%, 70.70% dan 71.52%. Gambar 5 adalah perbandingan *accuracy*, *precision*, *recall*, dan *f1-score* diantara kedua metode.



Gambar 5. Perbandingan Performa Model Klasifikasi

Nilai akurasi merupakan salah satu parameter penilaian terhadap metode klasifikasi. Nilai akurasi yang tinggi didapat ketika banyak data yang berhasil diklasifikasikan dengan benar sesuai dengan kelas sentimenya [28]. Sedangkan presisi berfungsi mengevaluasi kemampuan sistem untuk menemukan peringkat yang paling relevan, dan didefinisikan sebagai presentase dokumen yang diambil dan benar benar relevan terhadap query. Recall mengevaluasi kemampuan sistem untuk menemukan semua item yang relevan dari koleksi dokumen dan didefinisikan sebagai presentase dokumen yang relevan terhadap query [16]. Pada penelitian ini, metode Support Vector Machine cenderung lebih stabil karena Support Vector Machine menggunakan konsep klasifikasi dengan mencari hyperplane yang terbaik yaitu berfungsi sebagai pemisah dua kelas data dan mampu bekerja menggunakan kernel trik. Temuan ini sejalan dengan penelitian terdahulu [6]. Yang menyatakan bahwa kelebihan Support Vector Machine memiliki dimensi relatif yang tinggi sehingga dapat menggunakan fungsi kernel..

4. KESIMPULAN

Kesimpulan dari penelitian dan pembahasan adalah bahwa metode Naïve Bayes dan Support Vector Machine dapat diimplementasikan untuk klasifikasi sentimen terhadap vaksin Astrazeneca. Klasifikasi Naïve Bayes dapat digunakan untuk mengolah data dalam jumlah yang besar maupun kecil, Kekurangan dari metode Naïve Bayes dalam penelitian ini adalah Naïve Bayes bergantung pada kondisi dari masing-masing kejadian sehingga apabila kondisi nya bernilai nol maka probabilitas prediksi juga akan bernilai nol Hasil akurasi menggunakan Suport Vector Machine lebih besar dan akurat dari Naïve Bayes dalam mengklasifikasikan sentimen masyarakat terhadap vaksin Astrazeneca di media sosial Twitter. Metode Support Vector Machine menghasilkan akurasi sebesar 87.27% dari dari data uji dan untuk Naïve Bayes sebesar 76,81% Ini membuktikan metode SVM lebih akurat sebagai metode pengelompokkan untuk proses analisis sentimen masyarakat tentang vaksin Astrazeneca pada Twitter dibandingkan Naïve Bayes. Namun metode Support Vector Machine juga memiliki kelemahan yaitu jika data bersifat non-linear maka kemungkinan pengklasifikasian tidak memiliki generalitas yang tinggi.

REFERENCES

[1] S. Youse, R. Dara, S. Mubareka, and A. Papadopoulos, “International Journal of Infectious Diseases An analysis of COVID-19 vaccine sentiments and opinions on Twitter,” vol. 108, pp. 256–262, 2021, doi: 10.1016/j.ijid.2021.05.059.

[2] A. K. Napitupulu *et al.*, “ANALISIS KONSEP AL- ḌARŪRAH DALAM FATWA DSN -MUI ASTRAZENECA,” *At-Thullab J. Pendidik. Guru Madrasah Ibtidaiyah*, vol. 3, no. 14, pp. 748–767, 2021.

[3] L. Prasetyaning Widayanti and E. Kusumawati, “Hubungan Persepsi Tentang Efektifitas Vaksin Dengan Sikap Kesediaan Mengikuti Vaksinasi Covid-19,” *Hearty*, vol. 9, no. 2, p. 78, 2021, doi: 10.32832/hearty.v9i2.5400.

[4] K. RI, ITAGI, WHO, and UNICEF, “Survei Penerimaan Vaksin COVID-19 di Indonesia,” *Satuan Gugus Tugas Penanganan COVID-19*, no. November, pp. 1–26, 2020.

[5] N. Puri, E. A. Coomes, H. Haghbayan, and K. Gunaratne, “Social media and vaccine hesitancy : new updates for the era of COVID-19 and globalized infectious diseases,” *Hum. Vaccin. Immunother.*, vol. 16, no. 11, pp. 2586–2593, 2020, doi: 10.1080/21645515.2020.1780846.

[6] A. Saepulrohman, Sudin Saepudin, and D. Gustian, “Analisis Sentimen Kepuasan Pengguna Aplikasi Whatsapp Menggunakan Algoritma Naïve Bayes Dan Support Vector Machine,” *@is Best Account. Inf. Syst. Inf. Technol. Bus. Enterp.*, vol. 6, no. 2, pp. 91–105, 2021, [Online]. Available: https://rekayasa.nusaputra.ac.id/article/view/107%0Ahttps://rekayasa.nusaputra.ac.id/article/download/107/140.

[7] A. Mustopa, Hermanto, Anna, E. B. Pratama, A. Hendini, and D. Risdiansyah, “Analysis of user reviews for the pedulilindungi application on google play using the support vector machine and naive bayes algorithm based on particle swarm optimization,” *2020 5th Int. Conf. Informatics Comput. ICIC 2020*, vol. 2, 2020, doi: 10.1109/ICIC50835.2020.9288655.

Submitted: **08/06/2022**; Accepted: **05/07/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1545-1553
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4220

[8] R. Mahendrajaya, G. A. Buntoro, and M. B. Setyawan, "Analisis Sentimen Pengguna Gopay Menggunakan Metode Lexicon Based Dan Support Vector Machine," *Komputek*, vol. 3, no. 2, p. 52, 2019, doi: 10.24269/jkt.v3i2.270.
[9] T. N. Wijaya, Rini Indriati, and M. N. Muzaki, "Analisis Sentimen Opini Publik Tentang Undang- Undang Cipta Kerja Pada Twitter," *Jambura J. Electr. Electron. Eng.*, vol. 3, pp. 78–83, 2021.
[10] B. Laurensz and Eko Sediyono, "Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 10, no. 2, pp. 118–123, 2021, doi: 10.22146/jnteti.v10i2.1421.
[11] R. T. Aldisa, Azizah, and M. A. Abdullah, "Analisis Sentimen Mengenai Vaksin Sinovac di Media Sosial Twitter Menggunakan Metode Naïve bayes Classification," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 6, no. 3, pp. 1–5, 2022.
[12] M. R. A. Nasution and M. Hayaty, "Perbandingan Akurasi dan Waktu Proses Algoritma K-NN dan SVM dalam Analisis Sentimen Twitter," *J. Inform.*, vol. 6, no. 2, pp. 226–235, 2019, doi: 10.31311/ji.v6i2.5129.
[13] A. Suad A. and B. Wesam S., "Review of data preprocessing techniques in data mining.pdf," *J. Eng. Appl. Sci.*, vol. 12, no. 16, pp. 4102–4107, 2017, [Online]. Available: https://medwelljournals.com/abstract/?doi=jeasci.2017.4102.4107.
[14] F. Koto and G. Y. Rahmaningtyas, "InSet Lexicon : Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs InSet Lexicon : Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs," *IEEE*, no. December, pp. 391–393, 2017, doi: 10.1109/IALP.2017.8300625.
[15] D. Musfiroh *et al.*, "Analisis Sentimen terhadap Perkuliahan Daring di Indonesia dari Twitter Dataset Menggunakan InSet Lexicon," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 1, no. 1, pp. 24–33, 2021.
[16] R. Melita *et al.*, "PENERAPAN METODE TERM FREQUENCY INVERSE DOCUMENT FREQUENCY (TF-IDF) DAN COSINE SIMILARITY PADA SISTEM TEMU KEMBALI INFORMASI UNTUK MENGETAHUI SYARAH HADITS BERBASIS WEB (STUDI KASUS : SYARAH UMDATIL AHKAM)," *J. Tek. Inform.*, vol. 11, no. 2, 2018.
[17] T. Krisdiyanto, E. Maricha, and O. Nurharyanto, "Analisis Sentimen Opini Masyarakat Indonesia Terhadap Kebijakan PPKM pada Media Sosial Twitter Menggunakan Naïve Bayes Clasifiers," *CoreIT*, vol. 7, no. 1, pp. 32–37, 2021.
[18] H. C. Husada and A. S. Paramita, "Analisis Sentimen Pada Maskapai Penerbangan di Platform Twitter Menggunakan Algoritma Support Vector Machine (SVM) Sentiment Analysis of Airline on Twitter Platform Using Support Vector Machine (SVM) Algorithm," *IKADO*, vol. 10, no. 1, pp. 18–26, 2021, doi: 10.34148/teknika.v10i1.311.
[19] S. Yousefinaghani, R. Dara, S. Mubareka, A. Papadopoulos, and S. Sharif, "An analysis of COVID-19 vaccine sentiments and opinions on Twitter," *Int. J. Infect. Dis.*, vol. 108, pp. 256–262, 2021, doi: 10.1016/j.ijid.2021.05.059.
[20] T. Meisya *et al.*, "PERBANDINGAN KERNEL SUPPORT VECTOR MACHINE (SVM) DALAM PENERAPAN ANALISIS SENTIMEN VAKSINISASI COVID-19," *SINTECH*, vol. 4, no. 2, pp. 139–145, 2021.
[21] U. Verawardina, F. Edi, and R. Watrianthos, "Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes," *J. MEDIA Inform. BUDIDARMA*, vol. 5, pp. 157–163, 2021, doi: 10.30865/mib.v5i1.2604.
[22] C. D. Manning, P. Raghavan, and H. Schütze, *An Introduction to Information Retrieval*, no. c. 2009.
[23] G. N. Aulia and E. Patriya, "IMPLEMENTASI LEXICON BASED DAN NAIVE BAYES PADA ANALISIS SENTIMEN PENGGUNA TWITTER TOPIK PEMILIHAN PRESIDEN 2019," *J. Ilm. Inform. Kompute*, vol. 24, no. 2, pp. 140–153, 2019.
[24] Z. Alhaq, A. Mustopa, S. Mulyatun, and J. D. Santoso, "Penerapan Metode Support Vector Machine Untuk Analisis Sentimen Pengguna Twitter," *J. Inf. Syst. Manag.*, vol. 3, no. 2, pp. 44–49, 2021, doi: 10.24076/joism.2021v3i2.558.
[25] A. Sari, F. V., & Wibowo, "Analisis Sentimen Pelanggan Toko Online Jd. Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 2, no. 2, pp. 681–686, 2019.
[26] S. Utep and O. Kosheleva, "Why 70 / 30 or 80 / 20 Relation Between Training and Testing Sets : A Pedagogical Explanation Why 70 / 30 or 80 / 20 Relation Between Training and Testing Sets : A Pedagogical Explanation," 2018.
[27] N. Hardi, Y. Alkahfi, P. Handayani, W. Gata, and M. R. Firdaus, "Analisis Sentimen Physical Distancing pada Twitter Menggunakan Text Mining dengan Algoritma Naive Bayes Classifier," *Sistemasi*, vol. 10, no. 1, p. 131, 2021, doi: 10.32520/stmsi.v10i1.1118.
[28] F. Ratnawati, "Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter," *J. INOVTEK POLBENG*, vol. 3, no. 1, pp. 50–59, 2018.

Submitted: **08/06/2022**; Accepted: **05/07/2022**; Published: **25/07/2022**