Analisis Sentimen Masyarakat Indonesia Mengenai Vaksin COVID-19 pada Media Sosial Twitter Menggunakan Metode *Naïve Bayes Classifier* dan *Support Vector Machine*

Rizka Widya Permatasari dan Irhamah
Departemen Statistika, Institut Teknologi Sepuluh Nopember Surabaya (ITS)
e-mail: irhamah@statistika.its.ac.id

Abstrak*—World Health Organization (WHO) mendeklarasikan virus COVID-19 sebagai pandemi global pada 11 Maret 2020. Kondisi tersebut memberikan dampak langsung kepada seluruh masyarakat di dunia, dengan mulai diberlakukannya protokol ke-sehatan yang harus diterapkan pada seluruh aspek kegiatan, mulai dari pembatasan sosial hingga lockdown total yang menghambat seluruh kegiatan masyarakat. Salah satu cara yang dilakukan untuk mencegah penyebaran virus ini adalah dengan pemberian vaksin. Kegiatan vaksinasi mulai diberikan kepada masyarakat Indonesia pada bulan Januari 2021. Pada media sosial twitter, pro kontra vaksin COVID-19 sempat menjadi trending topic sehingga dirasa perlu untuk dilakukan penelitian tentang sentimen publik terhadap adanya kegiatan vaksinasi dalam memu-tus rantai penyebaran COVID-19 di Indonesia. Pada penelitian ini digunakan analisis klasifikasi teks yaitu Naïve Bayes Classifier (NBC) dan Support Vector Machine (SVM). NBC telah banyak digunakan dalam pe-nelitian mengenai Text Mining karena memiliki algoritma yang sederhana namun dapat menghasilkan akurasi yang tinggi, se-dangkan SVM memiliki kemampuan yang baik dalam mengolah data berdimensi besar dengan hasil yang efektif. Perbandingan kedua metode menggunakan 10 fold-stratified cross validation dengan kriteria kebaikan klasifikasi AUC dan akurasi menunjuk-kan bahwa SVM memiliki kinerja klasifikasi yang lebih baik dibanding NBC dan SVM kernel menghasilkan ketepatan klasifikasi lebih tinggi dibanding SVM kernel RBF.

***Kata Kunci*—AUC, COVID-19, Imbalanced, Naive Bayes Classifier, Stratified Cross Validation, Support Vector Machine, Twitter, Vaksin.**

I. PENDAHULUAN

VIRUS Corona atau *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2) adalah virus yang menyerang sistem pernapasan. Penyakit karena infeksi ini disebut dengan COVID-19, virus ini dapat menyebabkan gangguan ringan pada sistem pernapasaan, infeksi paru-paru yang berat hingga kematian. Virus ini menular dengan sangat cepat dan telah menyebar ke hampir semua negara, termasuk Indonesia, hanya dalam waktu beberapa bulan. World Health Organization (WHO) mendeklarasikan virus COVID-19 sebagai pandemi global pada 11 Maret 2020 [1].

Kondisi tersebut memberikan dampak langsung kepada seluruh masyarakat di dunia, dengan mulai diberlakukannya protokol kesehatan yang harus diterapkan pada seluruh aspek kegiatan, mulai dari pembatasan sosial hingga *lockdown* total yang menghambat seluruh kegiatan masyarakat. Pesatnya penyebaran COVID-19 dan bahaya yang akan muncul jika tidak segera ditangani, salah satu cara yang sangat mungkin untuk mencegah penyebaran virus ini adalah dengan mengembangkan vaksin [2]. Menyikapi hal tersebut, Pemerintah Indonesia juga turut aktif dalam rencana kegiatan vaksinasi yang akan diberikan kepada masyarakatnya. Pemerintah Indonesia resmi menandatangani formulir B vaksin GAVI Covax Facilities pada Kamis 7 Januari 2021, dengan penandatanganan tersebut, Indonesia akan memperoleh 108 juta dosis vaksin secara gratis. Kegiatan vaksinasi tersebut haruslah mempertimbangkan segala aspek terperinci agar rencana kegiatan vaksinasi dapat berjalan dengan baik dan terhindar dari hal-hal yang justru akan merugikan, mulai dari aspek kelayakan vaksin yang akan digunakan, resiko pasca pemakaian, sampai tahapan & prosedur dari pemberian vaksin hingga nantinya sampai ke masyarakat. Kegiatan vaksinasi tersebut juga haruslah mempertimbangkan berbagai masukan, diataranya adalah dengan melihat bagaimana respon dan opini masyarakat terhadap wacana vaksinasi COVID-19.

Pada media sosial *twitter*, vaksin COVID-19 sempat menjadi trending topic karena ramai dibahas oleh masyarakat Indonesia. Dengan fitur *thread* dan *trending*, *twitter* merupakan *micro blogging* yang cocok untuk dijadikan sebagai tempat berkumpul di dunia maya untuk curhat, bercerita, berdiskusi dan menyuarakan sebuah opini terhadap suatu topik. Opini yang disampaikan biasanya merupakan reaksi spontan dan emosional, yang bisa berupa opini positif maupun negative. Opini yang berada di twitter ini yang kemudian diubah menjadi data untuk dilakukan analisis sentimen.

Analisis sentimen merupakan salah satu teknik untuk mengekstrak sebuah informasi berupa sikap seseorang terhadap suatu isu atau kejadian dengan mengelompokkan polaritas dari sebuah teks [2]. Pengelompokkan tersebut dilakukan untuk melihat apakah teks tersebut bersifat positif, negatif atau netral. Analisis Sentimen dapat digunakan untuk mengetahui opini publik terhadap suatu isu seperti korupsi, dan demonstrasi berdasarkan data tekstual. Sebelum melakukan analisis sentiment, diperlukan *pre processing* teks dengan metode *text mining* untuk mengolah data teks agar siap untuk dianalisis. Terdapat banyak metode klasifikasi dalam ilmu statistika yang digunakan untuk klasifikasi teks diantaranya adalah *Naïve Bayes Classifier* (NBC*), K-Nearest Neigbour*, dan *Support Vector Machines* (SVM).

Penelitian ini akan menggunakan metode NBC dan SVM. Metode *Naïve Bayes Classifier* telah banyak digunakan

dalam penelitian mengenai *text mining,* hal itu disebabkan algoritma NBC yang sederhana namun memiliki akurasi yang tinggi [3]. Pemilihan metode SVM karena kemampuan generalisasi dalam mengklasifikasikan suatu *pattern* atau pola. Teknik ini berakar pada teori pembelajaran statistik dan telah menunjukkan hasil empiris yang menjanjikan dalam berbagai aplikasi praktis dari pengenalan digit tulisan tangan sampai kategorisasi teks. SVM juga bekerja sangat baik pada data dengan berbagai banyak dimensi dan menghindari kesulitan dari permasalahan dimensionalitas [4]. Berdasarkan penjelasan tersebut, maka permasalahan utama yang akan dibahas dalam penelitian ini adalah melakukan analisis sentiment pengguna *twitter* terhadap vaksin COVID-19 menggunakan NBC dan SVM

II. TINJAUAN PUSTAKA

A. Text Mining

Text mining merupakan salah satu bidang khusus dalam *data mining* yang digunakan untuk menganalisis suatu data berupa teks. *Text mining* adalah proses menganalisis suatu teks dengan menemukan pola-pola informasi dari teks yang tidak terstruktur untuk menggali informasi yang berkualitas dari suatu rangkaian teks yang terangkum dalam sebuah dokumen [5]. Pada dasarnya proses kerja dari *text mining* banyak mengapdopsi dari *data mining,* namun yang menjadi perbedaan adalah pola yang digunakan oleh *text mining* diambil dari sekumpulan bahasa alami yang tidak terstruktur sedangkan dalam *data mining* pola yang diambil dari *database* yang terstruktur. Penerapan *text mining* saat ini seperti kategorisasi teks, *text clustering,* ekstraksi konsep atau entitas, produksi taksonomi granular, *sentiment analysis,* penyimpulan dokumen, dan pemodelan relasi entitas.

B. Sentiment Analysis

Sentiment analysis atau *opinion mining* adalah bidang studi yang menganalisis pendapat, sentimen, evaluasi, sikap, penilaian dan emosi seseorang terhadap entitas seperti produk iklan, layanan, organisasi, individu, masalah, peristiwa, topik, ataupun kegiatan tertentu lainnya [2]. Tugas dasar dalam *sentiment analysis* adalah mengelompokkan teks yang ada dalam sebuah kalimat atau dokumen, kemudian menentukan pendapat yang dikemukakan apakah bersifat positif atau negatif.

C. Term Frequency Inverse Document Frequency (TF-IDF)

Term Frequency Inverse Document Frequency (*TF-IDF*) merupakan sebuah metode pembobotan yang dilakukan untuk ekstraksi data teks. Metode *TF-IDF* dilakukan dengan menghitung bobot dengan cara integrasi antara *term frequency* (*TF*) dan *inverse document frequency* (*IDF*). Rumus untuk menemukan pembobot dengan *TF-IDF* ditunjukkan pada Persamaan (1).

$$Z_y = TF_y \times IDF$$

$$IDF = \log\left(\frac{N}{DF_j}\right) \quad (1)$$

Keterangan :

Z_{ij} = bobot dari kata *i* pada artikel ke *j*,

N = jumlah seluruh *tweet*,

TF_{ij} = jumlah kemunculan kata *i* pada *tweet* ke-*j*,

DF_j = jumlah *tweet* ke-*j* yang mengandung kata *i*.

D. Stratified Cross Validation

Metode pembagian data menggunakan *K-fold cross validation* yang umum digunakan kurang sesuai apabila diterapkan pada masalah klasifikasi dengan ketidakseimbangan data (*imbalanced*). Hal itu disebabkan karena pembagian data menjadi K-*fold* memiliki distribusi probabilitas yang seragam sehingga menyebabkan satu atau lebih dari *fold* akan memiliki sedikit atau tidak memuat contoh dari kelas minoritas. *Stratified cross validation* merupakan teknik membagi data dengan memastikan bahwa dalam data *training* dan data *testing* harus ada perwakilan dari seluruh kelas yang ada dengan persentase yang sama [6]. *Stratified* dilakukan untuk memastikan bahwa dalam setiap fold merupakan representasi data yang baik.

E. Naïve Bayes Clasiffier (NBC)

Naïve Bayes Clasiffier merupakan salah satu teknik data *mining* yang sering digunakan untuk mengklasifikasikan data dalam jumlah yang besar dan dapat untuk mempresiksi probabilitas keanggotaan suatu *class* [3]. Kelebihan *Naïve Bayes Clasiffier* adalah algoritmnya sederhana tetapi memiliki akurasi yang tinggi. Klasifikasi *Naïve Bayes* yang merujuk pada teorema *bayes* mempunyai Persamaan (2).

$$P(B_k|Z_i, Z_2, \dots, Z_n) = \frac{P(Z_i, Z_2, \dots, Z_n|B_k)P(B_k)}{P(Z_i, Z_2, \dots, Z_n)} \quad (2)$$

Keterangan :

B_k = kelas data atau sentiment *tweet; k* = positif,negatif

Z_i = bobot *TF-IDF* pada kata kunci ke-*i; i* = 1,2,…,*n*

$P(B_k \mid Z_i)$ = probabilitas kategori *tweet* (B_k) berdasarkan kondisi bobot *TF-IDF* (Z_i) (*posterior probability*)

$P(Z_i \mid B_k)$ = peluang kondisi bobot TF-IDF (Z_i) berdasarkan kondisi kategori *tweet* (B_k) (*likelihood*)

$P(B_k)$ = peluang kategori *tweet* (*prior probability*)

$P(Z_{ij})$ = peluang bobot *TF-IDF* pada kata kunci (*evidence*).

Algoritma *Naïve Bayes Clasiffier* merupakan algoritma yang digunakan untuk mencari nilai probabilitas tertinggi untuk mengklasifikasi data *testing* pada kategori yang paling tepat. Berikut adalah algoritma dari pengklasifikasian *naïve bayes*.

1. Variabel Z_i merupakan sekumpulan dokumen yang direpresentasikan dengan pasangan atribut $Z_1, Z_2, .., Z_n$, dimana Z_1 adalah bobot *TF-IDF* pada kata kunci ke-1, Z_2 adalah bobot *TF-IDF* pada kata kunci ke-2. sedangkan B_k adalah himpunan kategori *tweet.*
2. Menghitung nilai $P(B_k)$ pada data *training* menggunakan Persamaan (3).

$$P(B_k) = \frac{|doc_k|}{|training|} \quad (3)$$

Keterangan :

$P(B_k)$ = peluang kategori tweet ke-*k* (*prior probability*)

Doc_k = jumlah *tweet* yang memiliki kategori *k*

training = jumlah *tweet* data *training.*

Untuk setiap probabilitas kata untuk setiap kategori dihitung pada saat *training* menggunakan Persamaan (4).

Tabel 1.
Fungsi Kernel Pada SVM

Fungsi Kernel	Rumus K (x_{m1},x_{m2})	Parameter
Linier	$\vec{x}'_{m_1}\vec{x}_{m_2}$	C
RBF	$\exp\left(-\frac{\left(\vec{x}_{m_1}-\vec{x}_{m_2}\right)^T\left(\vec{x}_{m_1}-\vec{x}_{m_2}\right)}{2\gamma^2}\right)$	γ dan C

$$P(Z_i|B_k)=\frac{n_i+1}{|n+kosakata|} \quad (4)$$

Keterangan :
n_i = jumlah kemunculan kata Z_i pada *tweet* berkategori B_k
n = jumlah seluruh kata dalam *tweet* dengan kategori B_k
kosakata = banyaknya kata dalam data *training*

3. Kemudian diklasifikasikan ke dalam kelompok yang memiliki probabilitas tertinggi menggunakan Persamaan (5).

$$V_{MAP}=\arg\max_{v_j=V} P(B_k)\prod_i P(Z_i|B_k) \quad (5)$$

F. Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah salah satu metode statistika yang mempelajari area yang memisahkan antar kategori dalam sebuah observasi. SVM adalah salah satu dari metode yang dikembangkan untuk mengatasi permasalahan yang tidak bisa diselesaikan dengan metode statistika klasik, terutama pada kasus klasifikasi dan prediksi [7]. SVM memiliki prinsip dasar *linier classifier* yaitu kasus klasifikasi yang secara *linier* dapat dipisahkan, namun SVM telah dikembangkan agar dapat bekerja pada problem non-*linier* dengan memasukkan konsep kernel pada *feature space* berdimensi tinggi.

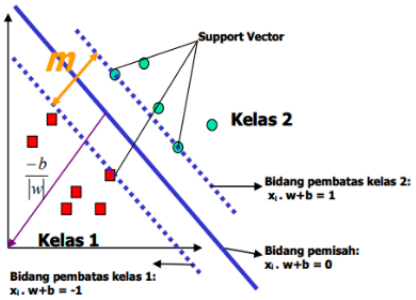
Konsep dari SVM pada *linearly separable* data adalah menemukan *hyperplane* yang optimum pada input *space*. Fungsi dari *hyperplane* itu digunakan sebagai pemisah dua buah kelas pada input *space* yang sering disimbolkan dengan -1 dan +1. Persamaan (6) merupakan Persamaan *hyperplane.*

$$\overrightarrow{x_m}.\vec{w}+b\geq+1 \text{ ; untuk } y_m=+1$$
$$\overrightarrow{x_m}.\vec{w}+b\leq-1 \text{ ; untuk } y_m=-1 \quad (6)$$

Keterangan :
w = normal bidang
x = himpunan data
y = label dari kelas x_m
b = posisi bidang relative terhadap pusat koordinat.

Untuk mendapatkan *hyperplane* optimum adalah dengan mencari *hyperplane* yang terletak ditengah-tengah antara dua bidang pembatas kelas (Gambar 1), yaitu dengan cara memaksimalkan margin atau jarak antara dua set objek dari kelas yang berbeda [8]. Nilai margin (jarak) antara bidang pembatas (berdasarkan rumus jarak garis ke titik pusat) adalah $\frac{2}{\|\vec{w}\|}$. Pencarian bidang pemisah terbaik dengan nilai margin terbesar dapat dirumuskan menjadi masalah optimasi konstrain (Persamaan (7)), yaitu dengan meminimalkan :

$$\min\frac{1}{2}\|\vec{w}\|^2 \quad (7)$$



Gambar 1. Bidang Pemisah Terbaik dengan Margin Terbesar.

dengan $y_m(\vec{x}_m.\vec{w}+b)-1\geq 0$ dan fungsi batasan $\sum_{m=1}^{M}\alpha_m\left[y_m(\vec{w}.\vec{x}_m+b)-1\right]$

Kemudian persamaan (7) diubah ke dalam formula *lagrangian* yang menggunakan *lagrange multiplier*.

$$\min_{w,b} L_p(\vec{w},b,\alpha)=\tfrac{1}{2}\|\vec{w}\|^2-\textstyle\sum_{m=1}^{M}\alpha_m y_m(\overrightarrow{x_m}.\vec{w}+b)+\sum_{m=1}^{M}\alpha_m \quad (8)$$

Dimana $\alpha_i\geq 0$ (nilai dari koefisien *lagrange*). Dengan meminimumkan *Lp* terhadap w dan b, maka dihasilkan persamaan (9),

$$\frac{\partial}{\partial b}L_p(\vec{w},b,\alpha)=0\Longrightarrow\textstyle\sum_{m=1}^{M}\alpha_m y_m=0$$
$$\frac{\partial}{\partial b}L_p(\vec{w},b,\alpha)=0\Longrightarrow w=\textstyle\sum_{m=1}^{M}\alpha_m y_m\overrightarrow{x_m} \quad (9)$$

Untuk menyederhanakan Persamaan (8) harus dilakukan transformasi ke dalam fungsi *Lagrange Multiplier* itu sendiri, sehingga persamaan (8) menjadi persamaan (10):

$$L_p(\vec{w},b,\alpha)=\tfrac{1}{2}\|\vec{w}\|^2-\textstyle\sum_{m=1}^{M}\alpha_m y_m(\overrightarrow{x_m}.\vec{w})-b\sum_{m=1}^{M}\alpha_m y_m+\sum_{m=1}^{M}\alpha_m \quad (10)$$

Berdasarkan persamaan (9), maka persamaan (10) menjadi persamaan (11) :

$$L_D(\alpha)=\textstyle\sum_{m=1}^{M}\alpha_m-\frac{1}{2}\sum_{m_1=1}^{M}\sum_{m_2=1}^{M}\alpha_{m_1}\alpha_{m_2}y_{m_1}y_{m_2}(\vec{x}_{m_2}^T\vec{x}_{m_2}) \quad (11)$$

Dan diperoleh *dual problem* yaitu pada persamaan (12)

$$\max_{\alpha} L_D(\alpha)=\textstyle\sum_{m=1}^{M}\alpha_m-\frac{1}{2}\sum_{m_1=1}^{M}\sum_{m_2=1}^{M}\alpha_{m_1}\alpha_{m_2}y_{m_1}y_{m_2}(\vec{x}_{m_2}^T\vec{x}_{m_2})$$
$$s.t\ \textstyle\sum_{m=1}^{M}\alpha_m y_m=0\ ;\ \alpha_m\geq 0 \quad (12)$$

Pada kasus linier *non-separable* beberapa data mungkin tidak bisa dikelompokkan secara benar atau terjadi *misclassification*. Persamaan (6) dimodifikasi dengan menambakan variabel *slack* ξ . H*yperplane* yang yang sudah dimodifikasi pada kasus *non-separable* ditulis dalam Persamaan (13) [8].

$$\overrightarrow{x_m}.\vec{w}+b\geq 1-\xi \text{ ; untuk kelas 1}$$
$$\overrightarrow{x_m}.\vec{w}+b\leq -1+\xi \text{ ; untuk kelas 2} \quad (13)$$

Formula pencarian bidang pemisah terbaik yaitu persamaan (7) berubah menjadi persamaan (14):

$$\min\tfrac{1}{2}\|\vec{w}\|^2+C(\textstyle\sum_{m=1}^{M}\xi_i) \quad (14)$$

SVM dapat bekerja pada data non-*linier* dengan menggunakan pendekatan kernel pada fitur data. Fungsi kernel yang digunakan untuk memetakan dimensi awal (dimensi yang lebih rendah) himpunan data ke dimensi baru

Tabel 3.
Confusion Matrix Dua Kelas

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

Tabel 4.
Struktur Data Analisis Sentimen

Tweet	Klasifikasi Sentimen	Kata Kunci (Z_1)	Kata Kunci (Z_2)	…	Kata Kunci (Z_i)
1	Positif	$Z_{1,1}$	$Z_{2,1}$	...	$Z_{i,1}$
2	Negatif	$Z_{1,2}$	$Z_{2,2}$	...	$Z_{i,2}$
3	Positif	$Z_{1,3}$	$Z_{2,3}$	...	$Z_{i,3}$
:	:	:	:	:	:
8225	Positif	$Z_{1,8225}$	$Z_{2,8225}$	...	$Z_{i,8225}$

Keterangan :
Z_{ij} = bobot *TF-IDF* pada kata kunci ke-*i* dan *tweet* ke-*j*
Y_i = Kelas data pada tweet ke-*j*

(dimensi yang relative lebih tinggi. Jika terdapat sebuah fungsi kernel K maka $K(x_{m_1} x_{m_2}) = \phi(x_{m_1})\phi(x_{m_2})$, sehingga fungsi yang dihasilkan dari *training* ditulis dalam Persamaan (15).

$$f(x_m) = \sum_{i=1}^{ns} a_i y_i K(\vec{x}_{m_1} \vec{x}_{m_2}) + b \quad (15)$$

Fungsi kernel yang umum digunakan ditampilkan pada Tabel 1.

G. Ketepatan Klasifikasi

Pengukuran ketepatan klasifikasi dilakukan untuk melihat performa klasifikasi yang telah dilakukan pada data *training* dan data *testing* dilakukan dengan menggunakan *confusion matrix* yang berisi informasi tentang kelas data asli yang direpresntasikan pada baris matriks dan kelas data hasil prediksi. Tabel 2 merupakan *confusion matrix* dari dua kelas data.

Pengukuran yang digunakan adalah akurasi, *specificity*, dan *sensitivity* dengan data yang *balanced* pada tiap kategorinya [3]. Persamaan (16) merupakan rumus dalam menghitung akurasi, *specificity*, dan *sensitivity*.

$$akurasi = \frac{TP+TN}{TN+TP+FN+FP} \quad (16)$$

Sedangkan pada data yang *imbalanced,* pengukuran ketepatan klasifikasi yang digunakan adalah AUC. Persamaan (17) adalah rumus untuk menghitung nilai AUC.

$$AUC = \frac{1}{2}\left(\frac{TP}{TP+FP} + \frac{TN}{TN+FP}\right) \quad (17)$$

III. METODOLOGI PENELITIAN

A. Sumber Data

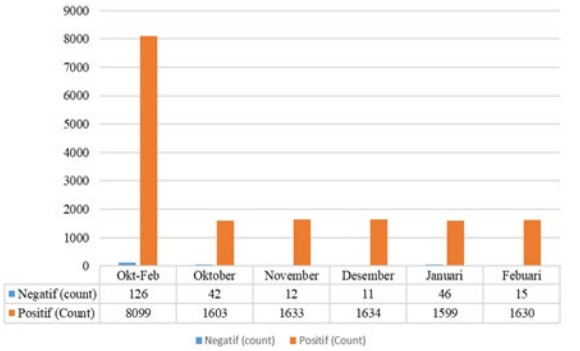
Data yang digunakan dalam penelitian ini adalah kumpulan *tweet* mengenai opini masyarakat mengenai vaksin COVID-19 dengan dua kata kunci yaitu "Vaksin Covid" dan "Vaksin Corona". Data *tweets* diambil pada tanggal 25 Oktober 2020 sampai 24 Febuari 2021 dari Twitter API (*Application Programming Interface*) sebanyak 8225 *tweet* karena ada keterbatasan dalam pengambilan data di Twitter.

B. Struktur Data

Struktur data yang digunakan dalam penelitian ini setelah dilakukan *pre processing* pada data teks *tweet* terdiri dari variabel prediktor yaitu kata dasar setiap *tweet* dan variabel respon yaitu klasifikasi sentiment *tweet* (positif dan negatif). Tabel 3 merupakan contoh struktur data penelitian untuk analisis sentimen.

Tabel 2.
Struktur Data Setelah *Pre Processing*

	…	Darurat	Vaksin	Covid	…	jokowi	harga	jangkau
Awal cabut izin darurat vaksin covid	...	1	1	1	…	0	0	0
Aman uji klinik covid henti	...	0	0	1	…	0	0	0
Beda vaksin vaksinasi imunisasi	...	0	1	0	…	0	0	0
Vaksin corona gratis bayar Jokowi harga jangkau	...	0	1	0	…	1	1	1



Gambar 2. *Bar Chart* Kategori Data Bulan Oktober 2020 - Febuari 2021.

C. Langkah Analisis

Langkah-langkah dalam menganalisis data penelitian ini adalah sebagai berikut.

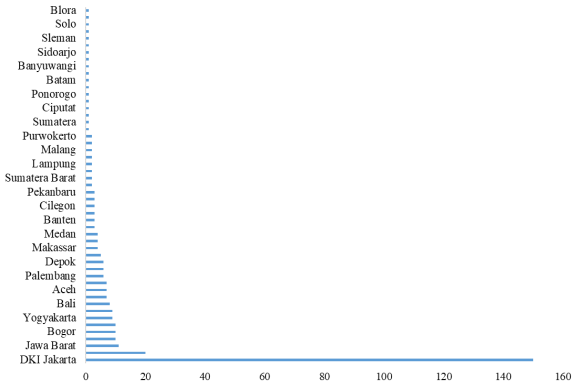
1. Mengambil data *tweet* dengan Twitter API.
2. Menyiapkan data *tweet*, daftar *stopwords*, dan kata dasar.
3. Memberikan label sentiment data *tweet* secara manual.
4. Melakukan *Pre-Processing Text.*
5. Melakukan pembobotan setiap kata atau *term weighting* menggunakan metode *TF-IDF.*
6. Klasifikasi data teks menggunakan metode NBC.
7. Klasifikasi data teks menggunakan SVM.
8. Membandingkan tingkat ketepatan klasifikasi dari model NBC dan SVM yang telah terbentuk.
9. Melakukan visualisasi dengan *wordcloud*.
10. Menarik kesimpulan dan saran.

IV. ANALISIS DAN PEMBAHASAN

A. Preprocessing dan Karakteristik Data

Data *tweet* mengenai vaksin COVID-19 pada media social *twitter* yang telah terkumpul dilakukan *preprocessing teks* meliputi *case folding, stopword, stemming* dan *tokenizing. Pre processing* data bertujuan untuk meningkatkan ketepatan klasifikasi dan mengurangi kesalahan klasifikasi data. Tabel 4 adalah struktur data yang telah dilakukan *pre processing teks*.

Data kategori sentiment sebagai variabel respon pada penelitian ini akan disajikan dalam *bar chart* untuk mengetahui perbandingan frekuensi antar bulan.



Gambar 3. *Bar Chart* Daerah Dengan Sentimen Negatif.

Tabel 6.
Probabilitas Klasifikasi NBC Data *Training*

Probabilitas Negatif	Probabilitas Positif	Keputusan
0.0985	0.9015	Positif
0.0026	0.9974	Positif
0.0007	0.9993	Positif
0.0007	0.9993	Positif
0.0007	0.9993	Positif
0.0448	0.9552	Positif
:	:	:
0.0001	0.9999	Positif
0.0001	0.9999	Positif
0.0004	0.9996	Positif
0.0004	0.9996	Positif

Tabel 7.
Hasil *Confusion Matrix* NBC

			Predict	
			Negatif	**Positif**
Training	Actual	Negatif	1	112
		Positif	1	7288
Testing	Actual	Negatif	0	13
		Positif	0	810

Gambar 2 menunjukkan bahwa pada bulan Oktober 2020 hingga Januari 2021 frekuensi antara kategori sentiment positif dan negative cenderung *imbalance*. Gambar 2 juga menunjukkan bahwa pada bulan Febuari 2021 merupakan bulan yang paling banyak mendapatkan sentimen positif dari public yaitu sebesar 98.91%.

Berdasarkan Gambar 2 dapat diketahui bahwa pada bulan oktober hingga febuari 2021 masih terdapat banyak masyarakat Indonesia yang beranggapan negative terhadap vaksin covid, oleh sebab itu perlu diketahui lebih lanjut daerah mana sajakah yang masih memberikan tanggapan negative terhadap vaksin COVID-19, berikut adalah *bar chart* daerah dengan sentiment negative pada bulan Oktober hingga Febuari 2021.

Gambar 3 menunjukkan bahwa mayoritas sentiment negative tentang vaksin COVID-19 berasal dari masyarakat di daerah DKI Jakarta, kemudian Jawa Barat dan Bogor.

B. Analisis Klasifikasi Menggunakan Metode Naïve Bayes Classifier (NBC)

Pada analisis klasifikasi data *tweet* mengenai vaksin COVID-19 dari bulan Oktober 2020 hingga Febuari 2021 metode yang digunakan adalah *Naïve Bayes Classifier.* Sebelum dilakukan analisis Langkah pertama yaitu membagi data menjadi data *training* dan data *testing* menggunakan *stratified 10-fold cross validation* dengan perbandingan 80:20.

Tabel 5.
Ketepatan Klasifikasi NBC

	Akurasi	**AUC**
Training	98.47	50.43
Testing	98.42	50.00

Tabel 8.
Hasil *Confusion Matrix* SVM *Linier*

			Predict	
			Negatif	Positif
Training	Actual	Negatif	64	50
		Positif	13	7276
Testing	Actual	Negatif	6	6
		Positif	9	801

Tabel 9.
Ketepatan Klasifikasi Terbaik dengan SVM *Linier*

	C	Akurasi	AUC
Training	100	99.15	77.98
Testing	100	98.18	70.33

Tabel 10.
Hasil *Confusion Matrix* SVM RBF

			Predict	
			Negatif	Positif
Training	Actual	Negatif	61	53
		Positif	10	7279
Testing	Actual	Negatif	6	6
		Positif	9	801

Klasifikasi dengan metode NBC menghasilkan probabilitas yang digunakan untuk menentukan apakah *tweet* masuk ke dalam kategori sentiment positif atau negative. Berikut adalah beberapa nilai probabilitas yang dihasilkan pada data *training*.

Nilai probabilitas *tweet*

Tabel 5 menunjukkan bahwa *tweet* mempunyai peluang untuk masuk ke dalam kategori sentimen sebesar nilai yang ada pada kedua kolom sentimen. Suatu *tweet* akan masuk ke dalam salah satu kategori sentimen apabila nilai probabilitasnya paling besar. Apabila probabilitas *tweet* masuk ke dalam kategori sentimen positif lebih besar dari probabilitas *tweet* masuk ke dalam kategori sentimen negatif maka *tweet* tersebut masuk ke dalam kategori sentimen positif dan sebaliknya. Dari probabilitas tersebut didapat kategori prediksi dari setiap tweet.

Selanjutnya pengukuran ketepatan klasifikasi dengan membentuk *confusion matrix.* Tabel 6 merupakan *confusion matrix* menggunakan *stratified* 10 *fold cross validation*.

Setelah terbentuk *confusion matrix*, maka langkah selanjutnya adalah melakukan perhitungan ketepatan klasifikasi.

Tabel 7**Error! Reference source not found.** menunjukkan bahwa pada klasifikasi data *training* diperoleh nilai AUC sebesar 50.43% dan pada data *testing* diperoleh nilai AUC sebesar 50.00%.

C. Analisis Klasifikasi Menggunakan Metode Support Vector Machine (SVM)

Klasifikasi menggunakan SVM sama seperti menggunakan metode NBC, perbedaannya hanya terletak di parameter. Pada metode SVM dengan kernel *radial basis function* (RBF) menggunakan dua parameter yaitu *C* dan γ, sedangkan kernel *linier* dilakukan dengan mempertimbangkan satu parameter yaitu *C*. Pada SVM dengan Kernel *linier* mempertimbangkan parameter *C* yang

Tabel 11.
Ketepatan Klasifikasi Terbaik dengan SVM RBF

	C	*Gamma*	Akurasi	AUC
Training	10000	0.01	99.15	76.68
Testing	10000	0.01	98.18	74.44

Tabel 82.
Perbandingan Ketepatan Klasifikasi

Data	AUC		
	NBC	SVM *linier*	SVM RBF
Training	50.43	77.98	76.78
Testing	50.00	74.44	74.44

akan dicoba yaitu 10^{-3} hingga 10^{4}. Tabel 8 merupakan *confusion matrix* menggunakan *stratified* 10 *fold cross validation*.

Tabel 9 adalah ketepatan klasifikasi dengan metode SVM kernel *linier*.

Klasifikasi dengan nilai *C* terbaik yaitu sebesar 100 dengan nilai ketepatan klasifikasi sebesar 77.98% pada data *training* data 70.33% pada data *testing*.

SVM dengan Kernel RBF mempertimbangkan parameter *C* dan γ. Parameter *C* yang akan dicoba yaitu 10^{-3} hingga 10^{4}, serta parameter γ yang akan dicoba yaitu 10^{-3} hingga 10^{4}. Tabel 10 merupakan *confusion matrix* menggunakan *stratified* 10 *fold cross validation*.

Setelah terbentuk *confusion matrix*, maka langkah selanjutnya adalah melakukan perhitungan ketepatan klasifikasi. Tabel 11 adalah ketepatan klasifikasi dengan metode SVM kernel RBF.

Nilai klasifikasi terbaik yang diperoleh pada parameter *C* dan γ yang paling optimum yaitu sebesar 10000 dan 0.01, dengan nilai AUC pada data *training* sebesar 76.68% dan data *testing* sebesar 74.44%.

D. Perbandingan Metode NBC dan SVM

Setelah mengetahui hasil masing-masing ketepatan klasifikasi pada kedua metode maka langkah selanjutnya adalah membandingkan hasil dari kedua metode tersebut. Tabel 12 merupakan hasil perbandingan metode *Naïve Bayes Classifier* dan *Support Vector Machine* berdasarkan nilai AUC dari bulan Oktober 2020 hingga Febuari 2021.

Tabel 12 menunjukkan bahwa secara keseluruhan performa metode SVM kernel *linier* lebih baik dibandingkan dengan metode NBC dan SVM kernel RBF, hal itu ditunjukkan dengan nilai AUC pada data *training* dan *testing* menghasilkan ketepatan klasifikasi yang lebih besar dibandingkan metode lainnya.

E. Model Support Vector Machine

Berdasarkan hasil analisis diperoleh kesimpulan bahwa SVM kernel Linier lebih baik dalam mengklasifikasikan sentiment masyarakat Indonesia mengenai Vaksinasi COVID-10. Hasil ketepatan klasifikasi terbaik pada bulan Oktober hingga Febuari 2021 pada metode SVM kernel Linier dengan nilai parameter *C* sebesar 100, kemudian dibentuk fungsi *hyperplane* dengan cara mensubtitusikan nilai *support vector* kategori positif pad x_{m1} dan nilai *support vector* kategori negative pada x_{m2}. Fungsi *hyperplane* dihitung dengan mensubtitusikan fungsi kernel pada Persamaan (18).

$$f(\vec{x}_m) = \sum_{m=1}^{656} \alpha_m y_m K\left(\vec{x}_{m_1} \vec{x}_{m_2}\right) + 1.00056837 \quad (18)$$

Gambar 4. *Wordcloud* Sentimen Negatif Masyarakat Terhadap Vaksin COVID-19.

Gambar 5. *Wordcloud* Sentimen Positif Masyarakat Terhadap Vaksin COVID-19.

Pada persamaan *hyperplane* nilai α_i merupakan vector koefisien atau *lagrange multiplier* dari *support vector*. Nilai yi merupakan label kategori yang memiliki dua nilai yaitu +1 untuk positif dan -1 untuk negative, serta *x* yang merupakan input yang akan diklasifikasikan.

F. Visualisasi Wordcloud

Visualisasi data teks menggunakan *wordcloud* digunakan untuk mengetahui kata-kata yang paling sering muncul pada data. Berikut adalah *wordcloud* mengenai sentiment negative masyarakat Indonesia terhadap vaksin COVID-19.

Gambar 4 menunjukkan kata-kata yang sering muncul pada sentiment negative yaitu kata “tidak”, ”tolak”, ”mati”, ”perintah”, ”nyawa”, ”rakyat”, ”bisnis” dan lain-lain. Berdasar-kan kata-kata yang sering muncul tersebut, menunjukkan bahwa public pengguna twitter banyak membahas bahwa mereka menolak menerima vaksin, vaksin hanya bisnis dari pemerintah Indonesia, vaksin mematikan, serta denda kepada masyarakat apabila menolak untuk di vaksinasi. Kata-kata lain yang mempunyai frekuensi yang jauh lebih kecil ditandai dengan ukuran *font* yang berbeda jauh seperti “flu”,”hiv”,”curiga” dan lain-lain. Gambar 4 adalah *wordcloud* mengenai sentiment positif masyarakat Indonesia terhadap vaksin COVID-19.

Gambar 5 adalah *wordcloud* mengenai sentiment positif masyarakat Indonesia terhadap vaksin COVID-19.

Gambar 5 menunjukkan kata-kata yang sering muncul pada sentiment positif yaitu kata “merah” “putih”,”strain”,”uji”, “klinis”,”kembang” dan “Indonesia”. Hal ini menunjukkan bahwa public pengguna Twitter banyak membahas mengenai Indonesia yang sedang mengembangkan vaksin merah putih, vaksin yang digunakan di Indonesia menggunakan strain yang sesuai, serta vaksin yang akan diberikan kepada masyarakat telah lolos uji klinis. Kata-kata lain yang mempunyai frekuensi

yang jauh lebih kecil ditandai dengan ukuran *font* yang berbeda jauh seperti “efek”,”samping”,”aman”,”efektif” dan lain-lain.

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan analisis data diperoleh kesimpulan yaitu nilai ketepatan klasifikasi menggunakan *Naïve Bayes Classifier* (NBC) pada data *training* dan *testing* diperoleh nilai AUC sebesar 50.43% dan 50%. Performa SVM kernel *linier* menghasilkan ketepatan klasifikasi yang lebih baik dibandingkan performa SVM kernel RBF. Hasil ketepatan klasifikasi menggunakan SVM kernel *linier* pada data *training* dan *testing* diperoleh nilai AUC sebesar 77.98% dan 74.44%. Secara keseluruhan perbandingan performa metode NBC dan SVM menunjukkan hasil bahwa performa metode SVM kernel *linier* lebih baik dalam mengklasifikasikan data teks sentiment masyarakat terhadap vaksin COVID-19. Kata-kata yang sering muncul pada sentiment negative yaitu kata “tidak”, ”tolak”, ”mati”, ”perintah”, ”nyawa”, ”rakyat”, ”bisnis” dan lain-lain. Serta kata-kata yang sering muncul pada sentiment positif yaitu kata “merah” “putih”, ”strain”, ”uji”, “klinis”, ”kembang” dan “Indonesia”.

B. Saran

Berdasarkan hasil analisis, rekomendasi yang dapat dipertimbangkan yaitu kepada pemerintah pusat dan pemerintah daerah khsuusnya daerah DKI Jakarta, Jawa Barat dan Bogor dengan jumlah masyarakat yang memiliki tanggapan negative terhadap vaksin COVID-19 paling banyak, sebaiknya lebih gencar dalam melakukan sosialisasi terkait pentingnya vaksin COVID-19 dalam memberantas penyebaran virus corona.

Saran untuk peneliti selanjutnya, penelitian serupa dapat dikembangkan dengan menggunakan media lain seperti Instagram, selain itu daftar kata pada *stopwards* dapat diupdate dengan daftar kata singkatan dan kata *slang* dalam bahasa Indonesia.

DAFTAR PUSTAKA

[1] C. Sohrabi *et al.*, “World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19),” *Elsevier Public Heal. Emerg. Collect.*, vol. 76, pp. 71–76, 2020, doi: 10.1016/j.ijsu.2020.02.034.

[2] B. Liu, *Sentiment Analysis and Opinion Mining*. California: Morgan & Claypool, 2012.

[3] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*. USA: Morgan Kaufmann, 2012.

[4] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Boston: Pearson Education, 2006.

[5] H. W. Ian, F. Eibe, and A. H. Mark, *Data Mining : Practical Machine Learning Tools and Techniques*. USA: Morgan Kaufmann, 2011.

[6] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” in *International Joint Conference on Artificial Intelligence (IJCAI)*, 1998, pp. 1137–11143.

[7] G. Williams, *Data Mining with Ratle and R : The Art of Excavating Data for Knowledge Discovery*. New York: Springer, 2011.

[8] E. E. Osuna, R. Freund, and F. Girosi, “Support Vector Machines: Training and Applications,” Massachusetts Institute of Technology Artificial Intelligence Laboratory, Cambridge, 1997.