Analisis Sentimen Mengenai Rencana Vaksinasi Covid-19 menggunakan Support Vector Machine dengan String Kernel

Devi Ayu Peramesti[1], Yuliant Sibaroni[2]

[1,2] Universitas Telkom, Bandung
[1]deviap@students.telkomuniversity.ac.id, [2]yuliant@telkomuniversity.ac.id

Abstrak

Pemerintah indonesia berencana melakukan vaksinasi masal sebagai upaya penanggulangan covid-19 dimana Indonesia memiliki kasus tertinggi di regional Asia Tenggara. Hal tersebut memicu berbagai opini masyarakat salah satunya di twitter. Vaksin covid-19 masih baru dan sedang dalam tahap uji coba pada manusia. Opini beragam ini dapat menjadi bahan masukan pemerintah dalam menyusun kebijakan vaksinasi masal. Untuk mengetahui gambaran opini diperlukan analisis sentimen. Yang nantinya diklasifikasikan menjadi kelas positif, kelas negatif dan kelas netral. Menggunakan metode *support vector machine* (SVM) dengan string kernel yang memiliki hasil uji terbaik. Hasil percobaan menunjukan masyarakat diprediksikan positif terkait kebijakan vaksinasi masal ini. Model terbaik untuk prediksi opini masyarakat didapat dengan menggunakan optimisasi parameter *gridsearch* dengan nilai performa f1-weighted sebesar 0.8373. Dengan menerapkan string kernel linear yang memiliki f1-score lebih tinggi dari kernel rbf,sigmoid,dan polynomial.

Kata kunci : analisis sentimen, SVM, kernel, multiclass,twitter,vaksinasi covid-19

Abstract

The Indonesian government plans to carry out mass vaccinations as an effort to deal with COVID-19, where Indonesia has the highest cases in the Southeast Asia region. This sparked various public opinions, one of which was on Twitter. The COVID-19 vaccine is still new and is currently in human trials. These diverse opinions can be used as input for the government in formulating mass vaccination policies. To find out the picture of opinion, sentiment analysis is needed. Which will be classified into positive class, negative class and neutral class. Using the support vector machine (SVM) method with a kernel string that has the best test results. The experimental results show that the community is predicted to be positively related to this mass vaccination policy. The best model for predicting public opinion is obtained by using gridsearch parameter optimization with an f1-weighted performance value of 0.8373. By implementing a linear kernel string that has a higher f1-score than the rbf, sigmoid, and polynomial kernels.

Keywords: sentiment analysis, SVM, kernel, multiclass,twitter,covid-19

1. Pendahuluan

COVID-19 adalah penyakit menular . WHO Oktober 2020, Indonesia peringkat pertama di Regional Asia Tenggara[1][2]. Berbagai upaya dilakukan salah satunya vaksinasi. Terdapat 100 kandidat vaksin COVID-19 yang sedang dikembangkan, dan dalam tahap uji coba pada manusia [3]. Kemenkes RI menuturkan syarat penerima vaksinasi COVID-19. Ada kelompok usia yang dikecualikan yakni kelompok usia 0-18 tahun, 60 tahun keatas serta orang dengan penyakit penyerta (komorbid) berat [4].

Rencana ini menuai banyak opini beragam di masyarakat salah satunya di twitter. Opini beragam tersebut, bisa digunakan oleh pemerintah sebagai masukan dalam menyusun kebijakan rencana vaksin masal covid-19. Berdasarkan riset pengguna media sosial Indonesia mencapai 56% dari total populasi[5]. Masuk peringkat 3 terbesar di Dunia untuk pengguna twitter[6] didominasi usia 18-34 tahun [7]. Dan rentang usianya masuk kedalam kelompok penerima vaksin.

Untuk mengetahui opini masyarakat diperlukan sentiment analisis. Bertujuan untuk menentukan sentimen orang-orang tentang suatu topik dengan menganalisis postingan mereka dan berbagai tindakan di media sosial [8]. Sentiment analisis bisa digunakan untuk menambang berbagai data di berbagai bidang ke ilmuwan. Misalnya, dalam bidang politik sentiment analisis pernah digunakan dalam memprediksi kemenangan pemilu di Indonesia pada tahun 2019 bersumber dari *tweet* dan *tag* terkait[9].

Mengenai rencana vaksinasi ini belum banyak penelitian sentiment yang membahas, terutama di Indonesia saat ini. Seperti pada penelitian yang membahasas pro kontra vaksinasi di Indonesia menggunakan metode LDA menunjukan penggunaan metode analisis yang lebih tervalidasi diharapkan karena penelitian ini terbatas pada penggunaan kamus positif-negatif [15].

Perbandingan mengenai keunggulan metode klasifikasi dapat dilihat dari penelitian menggunakan NBC, logistic regression, SVM, dan KNN. Hasilnya support vector machinea (SVM) terbukti lebih unggul dengan nilai *F-measure* sebesar 97.57% [11].Namun kelas klasifikasi sentimennya hanya menggunakan dua kelas sentimen yaitu positif dan negatif.

Keakuratan SVM bisa dilihat dari [12] hasil prediksi SVM berkorelasi positif dengan IEM (Iowa Electronics Markets), memprediksi Obama menangkan pemilu .Menyiratkan Twitter dapat dianggap sebagai sumber yang valid dalam memprediksi hasil. Kemudian, pemilihan fitur dan representasi yang tepat mempengaruhi peningkatan akurasi SVM salah satunya menggunakan pre-processing dan TF-IDF [14].

Berdasarkan klasifikasi berita menggunakan SVM dengan string kernel didapatkan RBF mampu menambahkan hasil akurasi menjadi 47,43% meski tidak terlalu signifikan. Penyebanya adalah preprocessing yang kurang. Kernel string adalah fungsi kernel seperti rbf , polynomial , linear dan sigmoid yang beroperasi pada string. Terdapat beberapa jenis kernel namun, pada penelitian digunakan dua kernel SVM yaitu RBF dan linear[20].

Untuk dapat mengetahui gambaran dari kebijakan rencana vaksinasi masal kita perlu mencari model klasifikasi dari algoritma SVM dengan kermel. Hal tersebut dapat memberikan solusi klasifikasi non-linear yang berpengaruh pada model[23]. Kelas klasifikasi terdiri dari positif,negatif dan netral. TF-IDF dan *preprocessing* dilakukan untuk peningkatan model. Kemudian prediksi opini masyarakat menggunakan model terbaik.

Kerangka penulisan penelitian ini adalah sebagai berikut: Studi terkait,Sistem penelitian yang dibangun , Evaluasi hasil penelitian dan juga menyimpulkan penulisan penelitian ini secara keseluruhan.

2. Studi Terkait

Metode analisis ini telah banyak digunakan pada big data sosial untuk mengumpulkan pendapat publik dari suatu subjek (layanan, produk, peristiwa, topik atau orang) di beberapa domain termasuk politik [3], pemasaran [4] dan kesehatan [7]. Analisis sentimen dapat dianalisi dengan *machine learning*[8]. Tweitter dianggap valid karena banyaknya penelitian yang bersumber dari twitter. Seperti pada penelitian mengenai prediksi pilpres Indonesia [9]. Kemudian penelitian yang memprediksi pemilu AS 2016 dengan twitter sebagai data sumbernya membangun kamus data berlabel kemudian, mengklasifikasikan tweet menjadi beberapa kelas[8].

Berdasarkan penelitian A.coralo menjelasakan metode analisis sentimen untuk menganalisis polaritas tweet yang memungkinkan pemerintah dapat menggambarkan pendapat masyarakat. Memberikan usulan pendekatan yang dapat dioptimalkan dengan menggunakan metode pembelajaran mesin untuk klasifikasi [17]. Didukung oleh penelitian R.Shahid(20) yang melakukan klasifikasi sentimen review produk menggunakan teknik Naïve Bayes dan Support Vector Machine. Karena machine learning dianggap lebih efektif dalam klasifikasi teks[10].

Dengan menggunakan algoritma klasifikasi Naïve Bayes, regresi logistik, mesin vektor dukungan, dan K-nearest neighbour. Hasilnya menunjukan support vector machine berkinerja lebih baik dibandingkan dengan 3 algoritma lain [11]. Selanjutnya pada penelitian ferdiana(2019) Pengujian dilakukan menggunakan algoritma SVM, KNN, dan SGD. Menunjukan SVM memiliki akurasi yang baik [14].

Meski klasifikasi menggunakan machine learning sangat popular tidak berarti hasil performansinya akan selalu baik. Oleh karena itu dibutuhkan pendekatan lain seperti melakukan proses *preprocessing.* Metode *preprocessing* teks yang sesuai termasuk transformasi dan pemfilteran data dapat secara signifikan meningkatkan kinerja pengklasifikasi [13]. Selain itu juga pembobotan TF-IDF ikut berperan dalam meningkatkan akurasi model klasifikasi [14].

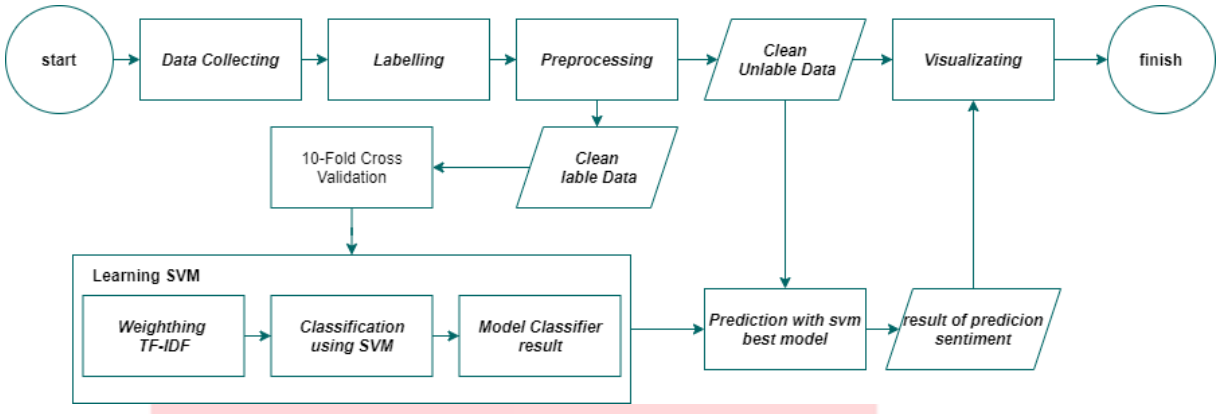
Terdapat penelitian serupa yaitu terkait vaksinasi covid – 19. Penelitian tersebut membahas mengenai pro dan kontra vaksin covid-19 dengan mengklasifikasikan menjadi kelas positif dan kelas negatif. Menggunakan metode LDA menghasilkan tanggapan masyarakat lebih positif terhadap wacana (30%). Model LDA yang dibangun juga dapat menangkap topik yang dibahas masyarakat terkait wacana vaksinasi[15].

Sedangkan pada penelitian ini akan dilakukan klasifikasi *multiclass* terdiri dari kelas positif, kelas negatif, dan kelas netral. Dan menggunakan algoritama SVM untuk melakukan klasifikasi. SVM mirip dengan LDA tetapi SVM dapat digunakan dalam meyelesaikan kasus linear atau non-linear dengan baik(*kernel trick*)[20][24]. Pemilihan fungsi kernel dan parameter mempengaruhi kinerja SVM. SVM mencapai hasil yang optimal jika parameter dapat diatur dengan baik salah satunya menggunakan optimasi parameter gridsearch[25].

3. Sistem yang Dibangun

3.1 Rancangan Sistem

Berikut merupakan rancangan sistem dalam penelitian ini:



Gambar 1. Diagram Rancangan Sistem Penelitian

3.2 Data Collecting

Datas collecting tweet dilakukan dengan menggunakan library tweepy [18]. Berikut timeline vaksinasi yang di publikasikan 14 januari 2021 [4][26][27:

Gambar 2. *Timeline* Vaskinasi Covid-19 Di Indonesia – News Liputan6.com

Berikut adalah kumpulan dari *hastag* yang digunakan:

Table 1. Daftar *Hastag Tweet* Untuk Proses *Data Collecting*

No	Hastag
1	Dukung Sehatkan Bangsa
2	Vaksin
3	VaksinYoVaksin
4	CeritaIndonesia2020
5	efek vaksin
6	Sinovac
7	takut vaksin
8	vaksin bahaya
9	vaksin corona
10	vaksin dimulai
11	vaksin gratis
12	vaksin indonesia
13	vaksinCovid19
14	vaksingratis
15	VaksinUntukNegeri
16	vaksin siap
17	vaksin sinopharm
18	vaksin untuk rakyat indonesia
19	vaksinasi
20	VaksinasiDimulai
21	vaksingratisuntukraktar
22	VaksinUntukKita

3.3 Pemberian Label

Tweet terkumpul sebanyak 70.000 menjadi 56.000 tweet karena duplikasi. Selanjutnya dibagi menjadi data berlabel dan tidak berlabel. Pelabelan dilakukan dengan proses manual sebanyak 5000 tweet.

Berikut adalah contoh dari data yang berlabel:

Table 2. Kelas Klasifikasi Tweet

Tweets	Label
Vaksin adalah konspirasi	Negatif
China aja nggak yakin vaksin bikinannya dan beli. Eh, INA malah beli buatan China…	Negatif
Target mencapai herd immunity akhir tahun ini dapat tercapai jika masyarakat Indonesia mendukung penuh program vaksinasi. #DosisVaksinAman https://t.co/1CvUTO9SyO	Positif
Jika masyarakat mendukung vaksinasi maka proses nya akan semakin lancar dan covid-19 juga akan perlahan menghilang #AyoVaksinTetap5M	Positif
Saya tidak ada menyebut lambatnya vaksinasi pemerintah karena kurangnya kapasitas Biofarma.	Netral
vaksin nanti disuntiknya di mana?	Netral

Pada proses pelabelan ini satu data akan dilabeli oleh tiga orang atau anotator. Berikut adalah Prosedur pelabelannya :

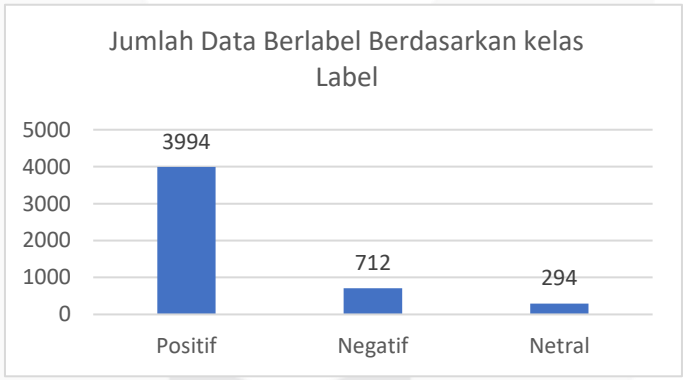
1) Annotator memberi label pada tweet bedasarkan opini pribadi.

2) Kemudian annotator akan mengkonversi opini dengan sebuah nilai berdasarkan kelas labelnya seperti berikut, jika tweet bersentimen:
   - positif menurut annotator maka nilai tweet tersebut adalah 1.
   - negative menurut annotator maka nilai tweet tersebut adalah -1.
   - netral menurut annotator maka nilai tweet tersebut adalah 0.
3) Akan dilakukan penjumlahan untuk nilai dari masing masing tweet. Dan didapatkan hasil berikut:
   - lebih dari 0 berlabel sentimen positif.
   - kurang dari 0 berlabel sentimen negatif.
   - sama dengan 0 berlabel sentimen netral.

Table 3. Proses Pelabelan Tweet

Tweet	Anotator			Hasil	Label
	1	2	3		
@jetravailaparis Sempet denger berita kayak gitu sih, tapi efeknya bakal sama gak ya sama Pfizer? Kayaknya tempo haâ€¦ https://t.co/ZcXxwvd3Pz	-1	0	1	0	Netral
udah pastinya jadi kebanggaan buat kita dong, udah 1juta nakes yang di vaksin, yeay sehat Indonesia https://t.co/LmNd54EMLo	1	1	1	3	Positif
Gue nggak tau gue boleh vaksin atau enggak. Belum ketemu dokter bedah syaraf lagi. Takut kerumah sakit euy /cupu/	0	-1	-1	-2	Negatif

Berikut adalah hasil dari pelabelan data :



Gambar 3. Jumlah Data Berlabel Berdasarkan Kelas Label

3.4 Pre Processing

3.4.1 Case Folding

Merubah huruf kapital menjadi huruf kecil karena akan dianggap berbeda.

3.4.2 Filtering

Pada proses filtering akan dilakukan pembersihan terhadap data tweet seperti menghapus string, gambar dan simbol. Pada penelitian ini untuk melakukan filtering akan menggunakan library RE [19].

3.4.3 Tokenization

Tahap tokenisasi merupakan tahap dimana dilakukan pemecahan kalimat menjadi kata-kata yang terpisah.[11].

3.4.4 Stemming

Stemming adalah proses mereduksi kata menjadi batang kata.[17]. Stemmer yang digunakan pada penelitian ini berasal dari library Sastrawi.

3.4.5 Stopword Removal

Stopword removal merupakan proses yang menghilanhkan kata – kata yang tidak memiliki arti. Proses *Stopword Removal* dilakukan berdasarkan kamus kata *stopword* yang ada pada library sastrawi.

Berikut adalah contoh dari masing – masing tahapan pre-processing :

Table 4. Hasil Tahapan Proses *Preprocessing*

Tweet Data	**Tahapan Preprocessing**	***Clean Tweet Data***
Bersama kita jaga tubuh dengan melakukan vaksin supaya terhindar dari corona virus menjaga daya tubuh penting https://t.co/bQl5KmLqs8	Case Folding	bersama kita jaga tubuh dengan melakukan vaksin supaya terhindar dari corona virus menjaga daya tubuh penting https://t.co/bQl5KmLqs8
	Filtering	bersama kita jaga tubuh dengan melakukan vaksin supaya terhindar dari corona virus menjaga daya tubuh penting
	Tokenization	Bersama, kita, jaga, tubuh, dengan, melakukan, vaksin, supaya, terhindar, dari, corona, virus, menjaga, daya, tubuh, penting
	Steamming	Bersama, kita, jaga, tubuh, dengan, melakukan, vaksin, supaya, hindar, dari, corona, virus, menjaga, daya, tubuh, penting
	Stopword Removal	Bersama, kita, jaga, tubuh, dengan, melakukan, vaksin, supaya, hindar, corona, virus, menjaga, daya, tubuh, penting

3.5 Pembobotan TF-IDF (Term Frequency - Inverse Document Frequency)

Proses *weighthing* atau pembobotan fitur *bag of words*. TF-IDF Mengkombinasikan dua faktor yaitu TF dan IDF.

$$tf_{t,d} = \frac{f_{t,d}}{\sum_{t' \in d} f_{t,d}} \quad (3.1)$$

TF merupakan frekuensi kemunculan kata biasnya ditransformasikan kedalam nilai log tf. Definisi konsep $tf_{t,d}$ memiliki keterangan sebagai berikut:

d = dokumen
t = term
$f_{t,d}$ = Jumlah kata pada suatu term untuk setiap dokumen
$\Sigma_{t' \in d} f_{t,d}$ = Jumalah dokumen yg menggandung term
$tf_{t,d}$ = *Term Frequency* (TF)

Denifisi Idf adalah sebagai berikut:

$$idf_i = log\left(\frac{N}{df_t}\right) \quad (3.2)$$

Memiliki keterangan sebagai berikut:

N = Jumlah dari seluruh dokumen
df_i = Jumlah kata pada term yang ada pada seluruh dokumen
$i\,df_i$= *Inverse Document Frquency*(IDF)

Dan TF-IDF memiliki persamaan sebagai berikut:

$$TF - IDF = tf_{t,d}\ x\ i\ df_i \tag{3.3}$$

Kata yang sering muncul seperti kata umum memiliki nilai idf yang rendah. TF-IDF melakukan ekstraksi sesuai dengan fitur yang memiliki bobot terbaik dalam suatu dokumen. Sehingga, semakin tinggi bobot suatu fitur dari dokumen validasi didalam dokumen data train maka dapat dipastikan bahwa dokumen validasi tersebut sudah pasti memiliki label yang sama dengan dokumen train yang memiliki bobot fitur tinggi tersebut. Berikut adalah ilustrasi dari pembobotan TF-IDF:

Table 5. Contoh Dokumen Kalimat

Dokumen	**Kalimat**
D1	vaksin penting untuk daya tahan tubuh kita
D2	kita vaksin agar terhindar virus

Table 6. Penghitungan TF-IDF

Kata	***Count***		df_i	$tf_{t,d}$ $\frac{f_{t,d}}{\sum_{t' \in d} f_{t,d}}$		$i\,df_i$ $log\left(\frac{N}{df_t}\right)$	**TF-IDF** $tf_{t,d}\ x\ i\ df_i$	
	D1	**D2**		D1	D1		D1	D2
vaksin	1	1	2	0.143	0.2	0	0	0
penting	1	0	1	0.143	0	0.301	0.043	0
untuk	1	0	1	0.143	0	0.301	0.043	0
daya	1	0	1	0.143	0	0.301	0.043	0
tahan	1	0	1	0.143	0	0.301	0.043	0
tubuh	1	0	1	0.143	0	0.301	0.043	0
kita	1	1	2	0.143	0.2	0	0	0
terhindar	0	1	1	0	0.2	0.301	0	0.060
agar	0	1	1	0	0.2	0.301	0	0.060
virus	0	1	1	0	0.2	0.301	0	0.060
$\sum_{t' \in d} f_{t,d}$	7	5						

3.6 Learning SVM

SVM bekerja dengan menaruh fitur bersumber dari data pelatihan pada sebuah bidang vektor. [14]. SVM bekerja menemuka hyperplane yang memisahkan data berhimpun menjadi dua kelas dalam bentuk linear [23].

Misalkan terdapat kelas -1 dan +1 dan dapat dipisahkan secara sempurna oleh hyperplane berdimensi d, yang di definisikan berikut :

$$w \cdot x + b = 0 \tag{3.4}$$

Dimana skalar b bisa bernilai negative, nol maupun positif.

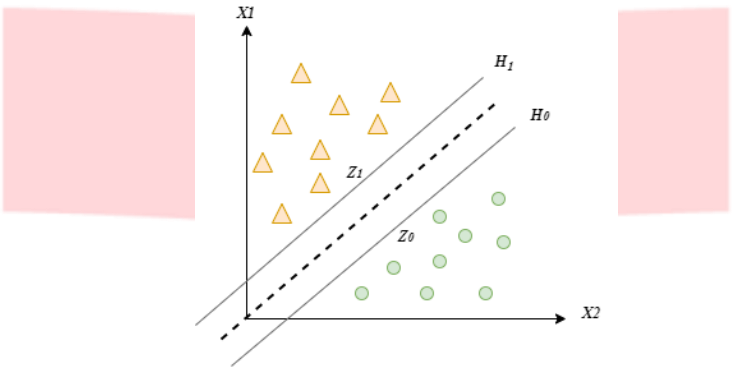
Misalkan H_0 adalah hyperplane memiliki persamaan :

$$w \cdot x + b = 1 \tag{3.5}$$

H_1 adalah hyperplane dengan persamaan :

$$w \cdot x + b = 0 \tag{3.6}$$

Kemudian xo adalah data yang berada pada hyperplane H0 dan z0 merupakan data yang ada di hyperplane H1.



Gambar 4. Contoh Pembagian Data Berdasarkan Kelas dengan *Hyperplane* SVM

Ilustrasinya dapat dilihat pada table berikut:

X1	X2	Kelas
-1	1	Negatif
-1	-1	Negatif
1	1	Positif

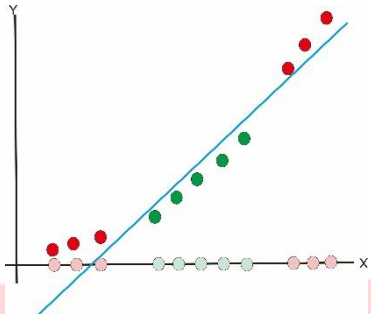
Gambar 5. Ilustrasi Kelas Hyperplane

Namun terdapat masalah yang tidak dapat ditangani secara linear(non linear). Masalah tersebut dapat diselesaikan dengan istilah *Kernel Trick*[20]. Yang mentrasformasikan data dari dimensi rendah ke dimensi tinggi. Kernel linear digunakan untuk memisahkan dengan satu garis. Kernel non linear memisahkan mampu memisahkan data dengan melengkungkan garis dan juga pada ruang dimensi yang tinggi[23]. Fungsi kernel yang umum digunakan yaitu:

Jenis Kernel	Definisi Rumus
Liniear	$K(x, x_k) = x_k^T x$
Polynomial	$K(x, x_k) = (x_k^T x + 1)^d$
Radial Basis Function	$K(x, x_k) = exp\{- \|\| x - x_2 \|\|_2^2 / \sigma^2\}$
Sigmoid	$K(x, x_k) = tanh[k\, x_k^T x + \theta]$

Gambar 6. Kernel Pada SVM dan pendefinisiannya

Fungsi kernel menghitung hubungan antara setiap pasangan titik, seolah berada pada dimensi lebih tinggi. Padahal tidak benar-benar melakukan trasnformasi. Terlepas dari bagaiamana hubungannya di hitung konsepnya tetap sama, seperti dijelaskan pada ilustrasi berikut:



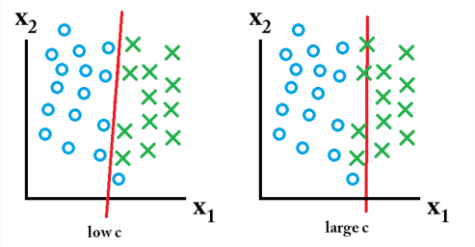
Gambar 7. Ilustrasi Cara Kerja Kernel SVM

Pada gambar ditas terlihat bahwa data memiliki 2 kategori tetapi sulit untuk pengklasifikasian linear memisahkan data tersebut dengan baik. Sehingga SVM bekerja dengan memindahkan data ke ruang dimensi tinggi. Dan kemudian menemukan dimensi Support Vector Classifier yang relative tinggi yang mampu mengkasifikasikan data secara efektif.

3.7 Parameter Optimization

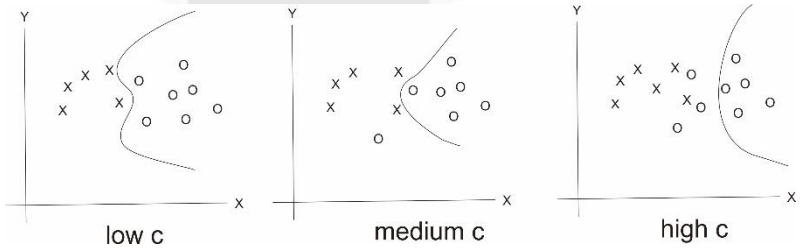
Parameter Optimization diperlukan pada SVM supaya membantu bangun model yang berakurasi tinggi. Pemilihan parameter yang tepat meningkatkan kinerja pembelajaran pada SVM. karenanya dibutuhkan parameter optimization untuk mendapatkan kombinasi parameter tepat. Salah satu metodenya adalah *GridSearch*. Yang menggunakan algoritma cross validation untuk melakukan serangkaian kombinasi pada parameter khususnya pada penelitian ini yang digunakan adalah parameter C dan Gamma.

C merupakan hyperparameter di SVM digunakan untuk mengontrol kesalahan atau margin seperti cotoh dibawah ini:



Gambar 8. Contoh Gambaran dari parameter C pada SVM

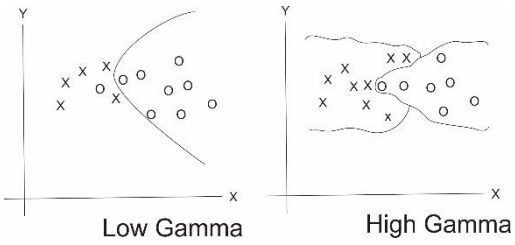
Pada gambar diatas ditunjukan bahwa low c memiliki tingkat kesalahan yang rendah sedangkan large c berarti memiliki kesalahan yang besar. Mari lihat contoh lainnya:



Gambar 9. Contoh Gambaran Parameter Pada SVM Dengan Kasus Data Berbeda

Kesalahan rendah tidak selalu memiliki model yang baik. Tetapi tergantung pada persebaran data. Terdiri dari berapa banyak data yang salah. Tidak ada aturan pasti bentuk *c* mana akan selalu berfungsi.

Gamma adalah hyperparameter yang menentukan berapa banyak lengkungan yang kita inginkan dalam batas keputusan nilai. *High Gamma* memiliki banyak lengkungan sedangakan *low gamma* lebih sedikit lengkungan. Seperti pada gambar berikut:



Gambar 10. Contoh Gambaran dari Parameter Gamma Pada SVM

Tidak ada nilai pasti C dan Gamma agar mendapatkan parameter terbaik hal tersebut sangat bergantung pada data. Sehingga jika range nilai luas dipercayai akan sangat berpeluang menemukan parameter terbaik[25]. Meski demikian terdapat hal penting yang harus diperhatikan dalam penentuan nilai dari C dan Gamma yaitu mengenai bias dan juga varian. Bias merupakan ketidakmampuan metode *machine learning* dalam menangkap hubungan data yang sebenarnya. Sedangkan Varian dalam istilah *machine learning* perbedaan fits diantara data set.

Nilai C yang besar memberikan bias yang rendah dan juga varians tinggi, bias rendah karena banyaknya kesalahan dalam klasifikasi. Sedangkan C yang kecil memberikan nilai bias yang tinggi dan varians yang rendah. Sedangkan nilai gamma yang kecil memberikan bias rendah dan varian tinggi. Hal tersebut berdampak pada tonjolan yang dihasilkan menjadi lebih runcing dan lebih jelas tidak melebar. Berdasarkan hal tersebut dan juga penelitian [25] berikut adalah daftar nilai dari C dan Gamma yang dipilih:

Table 7. Parameter yang akan digunakan

C	**Gamma**
0.1	1
1	0.1
10	0.01
100	0.001
1000	0.00001
	10

Untuk Linear kernel ada satu parameter penting yaitu C, pada polynomial ada 3 parameter penting yaitu C, Gamma dan Degree. Pada penelitian ini degree yang digunakan adalah 8. Kemudian pada kernel sigmoid dan RBF paremter penting ada dua yaitu C dan Gamma.

Pada Gambar dibawah ini terdapat ilustrasi gambar dari pencarian grid untuk parameter C dan Gamma. Untuk setiap pasangan data pada grid tersebut akan di gunakan pada setiap kernel yang digunakan yaitu rbf, svm , polynomial dan sigmoid. Himpunan pasangan data C dan Gamma yang memberikan nilai terbaik akan dipilih nilai tersebut sebagai nilai parameter model terbaik klasifikasi.

(C, Gamma)

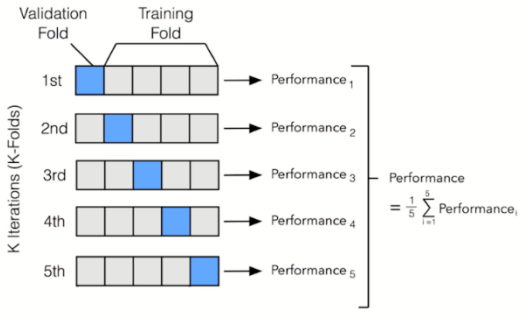
(0.1 , 1)	(1 , 1)	(10 , 1)	(100 , 1)	(1000 , 1)
(0.1 , 0.1)	(1 , 0.1)	(10 , 0.1)	(100 , 0.1)	(1000 , 0.1)
(0.1 , 0.01)	(1 , 0.01)	(10 , 0.01)	(100, 0.01)	(1000 , 0.01)
(0.1 , 0.001)	(1 , 0.001)	(10 , 0.001)	(100 , 0.001)	(1000, 0.001)
(0.1 , 0.00001)	(1 , 0.00001)	(10 , 0.00001)	(100 , 0.00001)	(1000 , 0.00001)
(0.1 , 10)	(1 , 10)	(10 , 10)	(100 , 10)	(1000 , 10)

Grid Search

Gambar 11. Ilustrasi Kombinasi dari Parameter C dan Gamma Menggunakan GridSearch

3.8 K-Fold Cross Validation

Untuk dapat mengukur kualitas model-model yang dibangun kita butuh metode ini. Selain itu juga metode ini Selain itu dapat membandingkan sejumlah metode klasifikasi,menyeleksi lalu memilih model mana yang terbaik. Untuk mendapatkan bias dan variansi lebih rendah umunya digunakan k=10.



Gambar 12. Contoh Gambaran *K-Fold Cross Validation*

himpunan data D dibagi dengan acak menjadi sub bagian yang disebut fold. Jika k =10 maka akan ada 10 data fold. Tiap fold berisi 1/k data. Kemudian setiap k- fold akan memiliki himpunan data dimana himpunannya sebanyak k-1 merukapan data uji sisanya merupakan data latih. Tidak akan ada data latih dan data train yang sama untuk seluruh fold [23].

3.9 Confusion Matrix Evaluation

Confusion Matrix Evaluation berguna untuk analisis kualitas model klasifikasi. Evaluasi tersebut dilakukan menggunakan suatu ukuran tertentu yaitu :

1) TP adalah true positives,jumlah tuple positif yang dilabeli dengan benar oleh model klasifikasi.
2) TN adala true negatives ,jumlah tuple negative yang dilabeli dengan bernar oleh model klasifikasi.
3) FP adalah false positeves,jumlah tuple negative yang salah dilabeli oleh model klasifikasi.
4) FN adalah false negatives, jumlah tuple yang salah dilabeli oleh model klasifikasi.

Berikut adalah ilustrasinya:

Table 8. Ukuran Performansi beserta definisinya

No	Ukuran	Definisi Rumus
1	Accuracy atau tingkat pengenalan	$\frac{TP + TN}{P + N}$

2	Recall atau sensitivity atau true positive rate	$\frac{TP}{P}$
3	Precision	$\frac{TP}{TP + FP}$
4	F atau F1 atau F-score atau rata-rata harmonic dari precision dan recall	$\frac{2\ x\ precision\ x\ recall}{Precission + Recall}$

Pada penelitian ini akan menggunakan nilai F1-Weighted sebagai nilai evaluasi hasil penelitian. F1-Weighted menghitung nilai F1-score dengan memperhatikan ketidakseimbangan jumlah data dari tiap kelas.

3.10 Prediksi Dan Klasifikasi Analisis Sentimen

Setelah mendapatkan model SVM mana yang akan digunakan untuk memprediksi opini masyarakat. Ditahap ini data tidak berlabel yang digunakan untuk prediksi sebanyak 51.000 tweet.

4. Eksperimen Dan Hasil

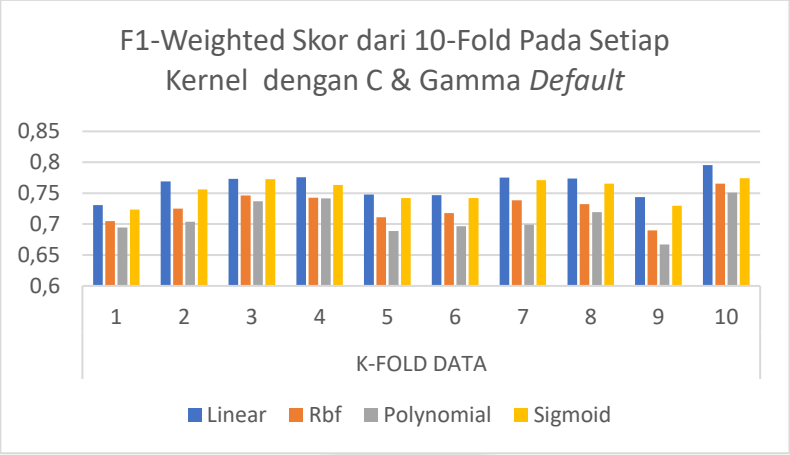
Bagian ini membahas hasil dan analisis dari pengujian penelitian. Selaras dengan tujuan penelitian yaitu mencari model klasifikasi terbaik berdasarkan string kernel pada SVM, prediksi sentimen masyrakat dengan melakukan klasifikasi data tidak berlabel, dan visualisasikan.

4.1 Hasil Pengujian

Berikut adalah skenario pengujian dari penelitian ini:

Skenario 1: Di skenario ini kita akan melakukan uji coba menggunakan fix parameter C dan Gamma *default*. Untuk mengetahui nilai uji dari kernel linear, rbf, polynomial dan sigmoid.

Berikut adalah Hasil pengujian dengan menggunakan fix parameter C dan Gamma *default*:



Gambar 13. Visualisasi hasil dari F1-Weighted pada setiap fold berasarkan kernel

Table 9. Hasil dari F1-Weighted Pada Setiap Data Fold Dengan Parameter Default Berdasarkan Kernel

Kernel	Fold Data Ke-										Averagre
	1	2	3	4	5	6	7	8	9	10	
Linear	0.731	0.769	0.773	0.776	0.748	0.747	0.775	0.774	0.744	0.796	0.763
Rbf	0.705	0.725	0.746	0.743	0.710	0.718	0.738	0.732	0.690	0.765	0.727
Polynomial	0.695	0.704	0.737	0.742	0.689	0.697	0.699	0.719	0.667	0.750	0.709
Sigmoid	0.723	0.756	0.773	0.763	0.742	0.742	0.771	0.766	0.730	0.774	0.754

Table 10. Model Terbaik dari Setiap Fold Data

Fold Data Ke-	Kernel	C	Gamma	Prediksi Data Test (f1-score)
0	linear	default	default	0.731
1	linear	default	default	0.769
2	linear	default	default	0.773
3	linear	default	default	0.776
4	linear	default	default	0.748
5	linear	default	default	0.747
6	linear	default	default	0.775
7	linear	default	default	0.774
8	linear	default	default	0.744
9	linear	default	default	0.796

Skenario 2: Di skenario ini kita akan melakukan dua uji coba menggunakan parameter optimization *GridSearch*. Untuk mengetahui nilai uji dari kernel linear, rbf, polynomial dan sigmoid.

Berikut adalah Hasil pengujian dengan menggunakan GridSearch pada data latih:

Table 11. Hasil Optimasi Parameter Terbaik Dari Setiap Data 10-Fold

Fold Data Ke-	Kernel	C	Gamma	Grid Search
0	rbf	10	0.1	0.832201
1	rbf	10	0.1	0.833477
2	rbf	10	0.1	0.829407
3	rbf	10	0.1	0.832042
4	rbf	10	0.1	0.828159
5	rbf	10	0.1	0.832587
6	rbf	10	0.1	0.830895
7	rbf	10	0.1	0.826762
8	rbf	1000	0.001	0.833478
9	rbf	10	0.1	0.831783

Setelah Mendapatkan parameter terbaik dari setiap data 10-fold dilakukan pengklasifikasian pada data uji. Berikut adalah hasil Klasifikasinya:

Table 12. Nilai F1-Weighted yang Didapatkan Dari Hasil Pengklasifikasian Data Uji Pada setiap Fold

Fold Data Ke-	Kernel	C	Gamma	Prediksi Data Test (f1-score)
0	rbf	10	0.1	0.815226
1	rbf	10	0.1	0.822539
2	rbf	10	0.1	0.860283
3	rbf	10	0.1	0.839993
4	rbf	10	0.1	0.842862
5	rbf	10	0.1	0.840658

6	rbf	10	0.1	0.83168
7	rbf	10	0.1	0.847982
8	rbf	1000	0.001	0.800195
9	rbf	10	0.1	0.845931

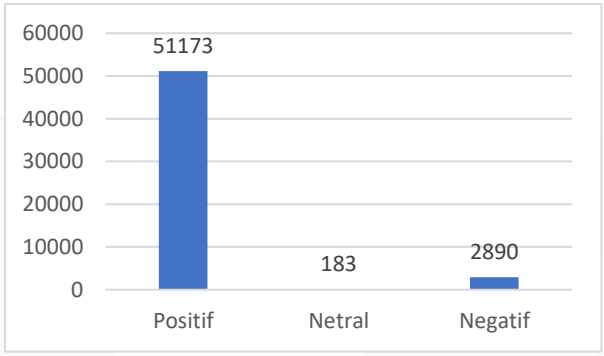
Dari dua skenario diatas kita dapat melihat bahwa dengan melakukan pengoptimalan parameter model klasifikasi menghasilkan nilai yang lebih baik. Selanjutnya kita melakukan pengoptimalan parameter dengan GridSearch pada seluruh data berlabel. Hasil Model yang didapatkan nantinya akan digunakan untuk melakukan prediksi pada data tidak berlabel. Berikut adalah hasil model yang didapatkan:

Table 13. Model Klasifikasi Terbaik Data Berlabel

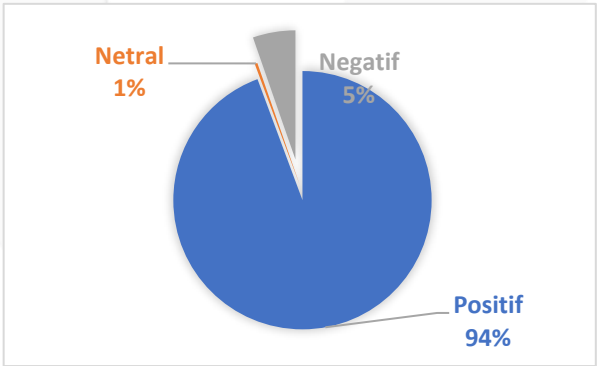
C	Gamma	Kernel	Grid Search
10	0.1	rbf	0.8373

4.2 Prediksi Sentimen Masyarakat

Setelah didapatkan model klasifikasi yang terbaik kita melakuakan prediksi sentimen masyrakat dengan menggunakan data tidak berlabel. Dan berikut adalah hasil dari prediksi sentimen.



Gambar 14. Hasil Prediksi Opini Masyarakat Berdasarkan Data Tidak Berlabel



Gambar 15. Persentasi Hasil Prediksi Opini Masyarakat Berdasarkan Data Tidak Berlabel

4.3 Analisis Hasil Pengujian

Hasil pengujian menunjukan masyarakat menyambut positif rencana vaksinasi masal covid -19. Terlihat selisih hasil klasifikasi positif, netral dan negatif sangatlah jauh. Disebabkan data label didominasi data positif. Oleh karena itu f1-score digunakan untuk dapat menghitung ketidakesimbangan data.

Kemudian dari hasil pengujian skenario 1 menggunakan fix parameter yang diaplikasikan pada masing-masing kernel menunjukan nilai yang cukup baik dan bervariasi pada setiap fold datanya. Hal tersebut dikarenakan

kita menggunakan komposisi pembagian data uji dan data latih berbeda beda dengan menggunakan metode k-fold cross validation. Hal tersebut memungkinkan kita untuk medapatkan skenario model terbaik. . Mengenai waktu yang dibutuhkan untuk menguji setiap kernel satu persatu dengan menggunakan fix parameter ini terbilang sangat lama. Oleh karena itu pada skenario tidak memungkinkan untuk melakukan lebih dari satu variasi parameter C dan Gamma. Berdasarkan rata-rata nilai setiap fold data, Pada skenario 1 ini kernel linear unggul dibanding kernel lainnya sebesar 0,763 . Hal ini dapat disebabkan oleh beberapa hal salah satu nya adalah kebutuhan parameter untuk setiap kernel itu berbeda. Kernel Linear berpeluang besar untuk mendapatkan nilai tertinggi karena kita hanya menggunakan fix parameter dari C dan Gamma saja. Selain itu dilihat dari setiap fold data kernel linear memiliki nilai lebih unggul dari kernel lainnya. Yang paling unggul ada pada data fol ke-10 dengan nilai f1-weighted sebesar 0,796 .

Pada pengujian di skenario 2 pengujiannya menggunakan optimisasi parameter *gridsearch* pada masing-masing kernel menunjukan nilai bervariasi pada setiap foldnya dan juga lebih unggul dari skenario 1. Hal tersebut tidak hanya dipengaruhi komposisi data saja tapi juga penggunaan nilai parameter C dan Gamma yang bervariasi. Hal ini berdampak pada luasnya eksperimen yang dapat dilakukan untuk mendapatkan model terbaik. Parameter bervariasi tersebut memungkinkan karena menggunakan *gridsearch*. Dan juga waktu yang dibutuhkan untuk memproses skenario ini jauh lebih cepat dibandingkan skenario 1. Pada skenario ini kita dapat melihat hasil nilai menggunakan *gridsearch* kernel rbf unggul untuk semua data fold. Hal tersebut mungkin karena parameter C dan gamma yang bervariasi sehingga setiap kernel memiliki kesempatan untuk mencari parameter terbaik. Dan dari 10 data dengan komposisi yang bervariasi ini kernel rbf dapat dengan baik menentukan nilai C dan Gamma yang sesuai sehingga hasil uji nya pun unggul. Data fold ke-2 memiliki nilai f1-weighted sebesar 0,860 .

5. Kesimpulan

Masuk dalam fenomena baru, masih sedikit yang melakukan penelitian sentiment analisis mengenai rencana vaksinasi covid-19. Oleh karena itu tugas akhir ini pun dapat dijadikan salah satu sumber dalam penelitian kasus kasus baru lainnya. Pada penelitian ini kita memprediksikan bahwa masyarakat beropini positif terkait kebijakan rencana vaksinasi masal covid-19. Dengan dilakukannya pengujian terhadap string kernel dari SVM didapatkan hasil bahwa string kernel linear unggul di skenario1 yang menggunakan *fix parameter* dengan nilai f1-weighted sebesar 0,796. Dan string kernel rbf unggul di skenario2 yang menggunakan optimisasi parameter *gridsearch* dengan nilai f1-weighted sebesar 0,860. Menunjukan pentingnya pemilihan parameter yang tepat sehingga bisa membantu proses *learning* SVM menghasilkan model yang terbaik. Selain itu kita dapat melihat seberapa berpengaruhnya komposisi dari data latih dan data uji untuk mendapatkan model klasifikasi yang terbaik. Untuk penelitian selanjutnya dapat menggunakan kelas label yang lebih spesifik dan bertingkat sehingga hasil prediksi dapat lebih akurat dan jelas. Kemudian penggunaan kombinasi jenis parameter lainnya dapat dipertimbangkan. Dengan adanya penelitian ini diharapkan dapat memberikan gambaran opini masyarakat mengenai kebijakan vaksinasi masal covid-19,sehingga dapat dijadikan masukan dalam menyusun kebijakan pemerintah khususnya dikondisi pandemi saat ini.

REFERENSI

[1] World Health Organization, “Weekly Operational Update on COVID-19 september 27, 2020,” *World Heal. Organ.*, no. October, pp. 1–10, 2020.

[2] CDC(Center for Deaseas Control and Prevention), “Long-Term Effects of COVID-19 | CDC.” https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects.html (accessed Dec. 03, 2020).

[3] World Health Organization, “Global Vaccine Action Plan,” *Vaccine*, vol. 31, pp. B5–B31, 2013, doi: 10.1016/j.vaccine.2013.02.015.

[4] 2020 Dirjen P2P Kemenkes, “Pemerintah Tengah Pastikan Keamanan dan Kehalalan Vaksin COVID-19 | Direktorat Jendral P2P.” http://p2p.kemkes.go.id/pemerintah-tengah-pastikan-keamanan-dan-kehalalan-vaksin-covid-19/ (accessed Nov. 30, 2020).

[5] Databooks, “Berapa Pengguna Media Sosial Indonesia? | Databoks.” https://databoks.katadata.co.id/datapublish/2019/02/08/berapa-pengguna-media-sosial-indonesia (accessed Dec. 03, 2020).

[6] statista, “• Indonesia: breakdown of social media users by age and gender 2020 | Statista.” https://www.statista.com/statistics/997297/indonesia-breakdown-social-media-users-age-gender/ (accessed Dec. 03, 2020).

[7] Databooks, “Indonesia Pengguna Twitter Terbesar Ketiga di Dunia | Databoks.” https://databoks.katadata.co.id/datapublish/2016/11/22/indonesia-pengguna-twitter-terbesar-ketiga-di-dunia (accessed Dec. 03, 2020).

[8] I. El Alaoui, Y. Gahi, R. Messoussi, Y. Chaabi, A. Todoskoff, and A. Kobi, “A novel adaptable approach for sentiment analysis on big social data,” *J. Big Data*, vol. 5, no. 1, 2018, doi:

10.1186/s40537-018-0120-0.
[9] W. Budiharto and M. Meiliana, "Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis," *J. Big Data*, vol. 5, no. 1, pp. 1–10, 2018, doi: 10.1186/s40537-018-0164-1.
[10] R. Shahid, S. T. Javed, and K. Zafar, "Feature selection based classification of sentiment analysis using Biogeography optimization algorithm," May 2017, doi: 10.1109/ICIEECT.2017.7916549.
[11] E. Sutoyo and A. Almaarif, "Twitter sentiment analysis of the relocation of Indonesia's capital city," *Bull. Electr. Eng. Informatics*, vol. 9, no. 4, pp. 1620–1630, 2020, doi: 10.11591/eei.v9i4.2352.
[12] A. Attarwala, S. Dimitrov, and A. Obeidi, "How efficient is Twitter: Predicting 2012 U.S. presidential elections using Support Vector Machine via Twitter and comparing against Iowa Electronic Markets," in *2017 Intelligent Systems Conference, IntelliSys 2017*, Mar. 2018, vol. 2018-January, pp. 646–652, doi: 10.1109/IntelliSys.2017.8324363.
[13] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," *Procedia Comput. Sci.*, vol. 17, pp. 26–32, 2013, doi: 10.1016/j.procs.2013.05.005.
[14] R. Ferdiana, F. Jatmiko, D. D. Purwanti, A. S. T. Ayu, and W. F. Dicka, "Dataset Indonesia untuk Analisis Sentimen," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 8, no. 4, p. 334, 2019, doi: 10.22146/jnteti.v8i4.533.
[15] F. F. Rachman and S. Pramana, "Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter," *Heal. Inf. Manag. J. ISSN*, vol. 8, no. 2, pp. 2655–9129, 2020, [Online]. Available: https://inohim.esaunggul.ac.id/index.php/INO/article/view/223.
[16] P. Singh, R. S. Sawhney, and K. S. Kahlon, "Sentiment analysis of demonetization of 500 & 1000 rupee banknotes by Indian government," *ICT Express*, vol. 4, no. 3, pp. 124–129, 2018, doi: 10.1016/j.icte.2017.03.001.
[17] A. Corallo *et al.*, "Sentiment analysis for government: An optimized approach," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9166, no. July, pp. 98–112, 2015, doi: 10.1007/978-3-319-21024-7_7.
[18] tweepy, "API Reference — tweepy 3.9.0 documentation." http://docs.tweepy.org/en/latest/api.html#id4 (accessed Dec. 03, 2020).
[19] Python, "Regular Expression HOWTO — Python 3.3.7 documentation." https://docs.python.org/3.3/howto/regex.html (accessed Dec. 04, 2020).
[20] Honakan, Adiwijaya, and S. AL-Faraby, "Analisis Dan Implementasi Support Vector Machine Dengan String Kernel Dalam Melakukan Klasifikasi Berita Berbahasa Indonesia Analysis and Implementation Support Vector Machine With String Kernel for Classification indonesian news," vol. 5, no. 1, pp. 1701–1710, 2018.
[21] Skicit-learn, "sklearn.svm.SVC — scikit-learn 0.23.2 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#examples-using-sklearn-svm-svc (accessed Dec. 04, 2020).
[22] Skicit-learn, "1.4. Support Vector Machines — scikit-learn 0.23.2 documentation." https://scikit-learn.org/stable/modules/svm.html (accessed Dec. 04, 2020).
[23] Suyanto, *machine Learning Tingkat Dasar dan Lanjut*. Bandung: Informatika Bandung, 2018.
[24] M. T. . Damarsasi Wilogo1, Erwin Budi Setiawan, S.Si., M.T. 2, Yuliant Sibaroni, S.Si., "Mendeteksi Spammers di Twitter dengan SVM Classifier," vol. 5, no. 3, pp. 8249–8258, 2018.
[25] I. Syarif, A. Prugel-Bennett, and G. Wills, "SVM Parameter Optimization using Grid Search and Genetic Algorithm to Improve Classification Performance," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 14, no. 4, p. 1502, 2016, doi: 10.12928/telkomnika.v14i4.3956.
[26] "Menkes: Masalah Laju Vaksinasi Bukan di Lokasi, tapi di Ketersediaan Vaksin COVID-19 - Health Liputan6.com." https://www.liputan6.com/health/read/4506556/menkes-masalah-laju-vaksinasi-bukan-di-lokasi-tapi-di-ketersediaan-vaksin-covid-19 (accessed Sep. 19, 2021).
[27] "Dimulainya Vaksinasi Covid-19 di Indonesia." https://nasional.kompas.com/read/2021/01/14/06572221/dimulainya-vaksinasi-covid-19-di-indonesia (accessed Sep. 19, 2021).