JAMBURA JOURNAL OF PROBABILITY AND STATISTICS
Volume 3 Nomor 2, November 2022
Journal Homepage: http://ejurnal.ung.ac.id/index.php/jps
DOI: https://doi.org/10.34312/jjps.v3i2.16663

ANALISIS SENTIMEN MASYARAKAT PADA KEBIJAKAN VAKSINASI COVID-19 DI TWITTER MENGGUNAKAN METODE MESIN VEKTOR PENDUKUNG DENGAN KERNEL *RADIAL BASIS FUNCTION* BERBASIS FITUR LEKSIKON

Sri Mulyani [1], Sri Astuti Thamrin[2,*], Siswanto [3]
[1,2,3] Departemen Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin

[*]**e-mail**: tuti@unhas.ac.id

Abstrak
Twitter merupakan salah satu platform media sosial yang memungkinkan orang untuk berkomentar, berbalas komentar dan mengetahui berbagai hal yang tengah menjadi *trending*. Melalui *twitter* ini, masyarakat berkomentar terkait pandemi *Coronavirus* disease 2019 (COVID-19) dan kebijakan vaksinasi yang pemerintah keluarkan untuk melindungi masyarakat dari COVID-19. Untuk menganalisis berbagai komentar masyarakat di twitter pada kebijakan vaksinasi COVID-19 ini digunakan analisis sentimen. Pada analisis sentimen, informasi berupa data teks dapat diekstrak untuk memperoleh pengetahuan (*knowledge*), yang dikenal dengan *text mining*. Studi ini bertujuan untuk mengklasifikasi komentar masyarakat terkait kebijakan vaksinasi COVID-19 di Indonesia di twitter dan membandingkan kinerja metode dari hasil klasifikasi itu. Metode yang digunakan dalam studi ini adalah mesin vektor pendukung berbasis fitur leksikon dengan kernel *radial basis function* (RBF) dan tanpa fitur leksikon. Sebanyak 2981 *tweet* diekstrak dari twitter. Hasil studi ini menunjukkan kinerja mesin vektor pendukung dengan RBF tanpa fitur leksikon dalam klasifikasi komentar masyarakat terhadap kebijakan vaksinasi tersebut diperoleh nilai akurasi, *G-mean* dan AUC masing-masing 83%, 50% dan 61,35%. Sementara, kinerja metode tersebut dengan fitur berbasis leksikon diperoleh klasifikasi komentar masyarakat pada vaksinasi tersebut dengan nilai akurasi, *G-mean* dan AUC masing-masing sebesar 90%, 86,63% dan 87%. Hal ini menunjukkan bahwa metode mesin vektor pendukung dengan kernel RBF berbasis fitur leksikon memberikan hasil klasifikasi komentar masyarakat pada vaksinasi COVID-19 yang lebih baik dibandingkan dengan mesin vektor pendukung tanpa fitur leksikon.
Kata Kunci: Analisis Sentimen, COVID-19, Leksikon, Mesin Vektor Pendukung, Radial Basis Function, Vaksinasi

Abstract
Twitter is a social media platform that allows people to comment, reply to comments and find out what's trending. Through this twitter, the public commented on the Corona virus disease 2019 (COVID-19) pandemic and the vaccination policy that the government issued to protect the public from COVID-19. To analyze various public comments on Twitter on the COVID-19 vaccination policy, sentiment analysis was used. In sentiment analysis, information in the form of text data can be extracted to obtain knowledge, which is known as text mining. This study aims to classify public comments related to the COVID-19 vaccination policy in Indonesia on twitter and compare the performance of the method from the results of that classification. The method used in this study is a lexicon feature-based support vector engine with a radial basis function (RBF) kernel and without lexicon features. A total of 2981 tweets were extracted from twitter. The results of this study show that the performance of the supporting vector machine with RBF without lexicon features in the classification of public comments on the vaccination policy obtained accuracy values, G-mean and AUC of 83%, 50% and 61.35%, respectively. Meanwhile, the performance of this method with lexicon-based features obtained the classification of public comments on the vaccination with accuracy, G-mean and AUC values of 90%, 86.63% and 87%, respectively. This shows that the supporting vector machine method with the RBF kernel based on the lexicon feature

Received: 2022-09-28 | Accepted: 2022-11-06 | Available Online: 2022-11-30

provides better results for classifying public comments on COVID-19 vaccination compared to the supporting vector machine without the lexicon feature.
***Keywords**: Sentiment Analysis, Lexicon, COVID-19, Support Vector Machine, Radial Basis Function, Vaccination*

1. PENDAHULUAN

Media sosial adalah platform digital yang memberikan fasilitas kepada para penggunanya untuk saling berkomunikasi, sehingga menghasilkan data internet yang sangat banyak (Dhawan and Zanini, 2014). Twitter merupakan salah satu media sosial yang populer. Pada *twitter*, berbagai masalah dapat dikemukakan oleh pengguna (Kurniawan, 2017). Untuk memesan dan mengatur informasi di *twitter*, tanpa batasan waktu, dapat digunakan fasilitas *library* di *python*, yang dinamakan *Snscrape.* Dimasa pandemi COVID-19, pemerintah mengeluarkan kebijakan vaksinasi untuk melindungi masyarakat dari COVID-19. Pada awal digulirkannya kebijakan ini, berbagai macam komentar yang disampaikan masyarakat di media *twitter*.

Informasi terkait vaksinasi COVID-19 yang beredar di *twitter* ini merupakan sumber data berbentuk teks yang dapat diekstrak. Menganalisis data dalam bentuk teks-teks ini memerlukan teknik analisis tertentu. Teks-teks itu berisi komentar pro dan kontra masyarakat pada kebijakan vaksinasi COVID-19. Untuk mengolah dan menganalisis informasi tersebut agar menghasilkan suatu pengetahuan maka, diperlukan suatu metode tertentu. *Text mining* merupakan salah satu pendekatan yang umum digunakan untuk mengetahui tanggapan masyarakat di *twitter*, termasuk tanggapan pada vaksinasi COVID-19. Melalui *text mining* dapat dianalisis sentimen masyarakat pada vaksinasi COVID-19 (Liu, 2012). Namun, dalam kenyataannya, hanya polaritas kalimat yang diperhatikan dan mengabaikan nilai kata yang ada di tweet pada saat melakukan analisis sentimen. Untuk mengatasi masalah itu, fitur berbasis leksikon dapat digunakan dengan cara memberi bobot berdasarkan kamus atau *lexical*. Selanjutnya, hal itu diolah untuk menghasilkan informasi sentimen berbahasa Indonesia.

Komentar masyarakat pada vaksinasi COVID-19 dapat diklasifikasikan menjadi pro dan kontra. Metode mesin vektor pendukung merupakan salah satu metode yang umum digunakan untuk klasifikasi (Nugroho *et al.*, 2003; Vidya, 2015; Pisner and Schnyer, 2020). Konsep dasar metode ini adalah dikenal dengan nama *linear classifier*. Dalam perkembangannya metode ini selanjutnya dikembangkan pada problem non-linear dengan menggunakan *kernel trick* di ruang berdimensi tinggi (Nugroho *et al.*, 2003). Pada penelitian Vidya *et al.* (2015) tentang analisis sentimen, dibandingkan dengan kinerja naïve Bayes dan pohon keputusan, akurasi metode mesin vektor pendukung lebih tinggi. Berdasarkan hal itu, pada studi ini akan dianalisis komentar masyarakat pada kebijakan vaksinasi COVID-19 di *twitter* menggunakan metode mesin vektor pendukung baik dengan fitur leksikon maupun tidak. Selain itu, masing-masing metode kinerja juga akan dibandingkan kinerjanya pada studi ini.

2. METODE PENELITIAN

Pelaksanaan vaksinasi tahap pertama yang dicanangkan pemerintah berlangsung dalam dua periode, yakni periode 1 berlangsung dari Januari hingga April 2021, dan periode 2 berlangsung dari April 2021 hingga Maret 2022. Studi ini menggunakan data kumpulan *tweet* yang diambil dari media *twitter* pada periode waktu 1 Januari sampai 9 April 2021. Periode pengambilan data dilatarbelakangi karena rentang waktu tersebut merupakan periode pertama yang menimbulkan banyak pro dan kontra di masyarakat. Kata kunci yang digunakan adalah "Vaksinasi COVID-19". Data diperoleh dari sncrape sebanyak 10.000 data

tweet dan format data yang diambil berbentuk teks. Variabel yang digunakan terdiri dari variabel prediktor berupa kata dasar pada setiap tweet dan variabel respon berupa sentimen positif dan negatif. Data hasil *tweet* dibagi menjadi data latih (80%) dan data uji (20%) dan menggunakan 5-*fold cross validation* untuk mencatat nilai evaluasi kinerja dari model dengan menggunakan *confusion matrix.* Dengan menggunakan 5-*fold cross validation*, diharapkan akan menghasilkan nilai akurasi yang baik untuk kasus ini.

2.1 Pra-proses Teks

Pra-proses Teks (*Text pre-processing*) adalah tahap pertama dari pemrosesan teks. Selama fase ini, dokumen secara opsional diubah menjadi data terstruktur untuk diproses lebih lanjut dalam proses penambangan teks. Hal ini bertujuan agar akurasi klasifikasi data dapat ditingkatkan (Kurniawan, 2017). Adapun tahapan pada *text pre-processing* yaitu *case folding, data cleansing, stemming, stopword,* dan *tokenizing*.

2.2 *Term Frequency Inverse Document Frequency*

Term Frequency-Inverse Document Frequency (TF-IDF) merupakan proses pembobotan yang dilakukan dengan mengubah data tekstual ke dalam data numerik pada tiap kata atau fitur. Evaluasi pentingnya sebuah kata di dalam sebuah dokumen dilakukan pada tahapan proses TF-IDF. TF (*Term Frequency)* menunjukkan seberapa penting sebuah kata di dalam tiap dokumen tersebut, dan DF (*Document Frequency)* menunjukkan seberapa umum kata tersebut digunakan. Kebalikan dari nilai DF-IDF adalah *index document frequency*. Bobot kata kemudian diperoleh dari hasil perkalian antara TF dan IDF pada proses TF-IDF. Semakin sering kata muncul dalam dokumen, maka bobot kata semakin besar dan sebaliknya, bobot kata akan semakin kecil jika muncul dalam banyak dokumen (Septian *et al.*, 2019). Perhitungan bobot kata TF-IDF diberikan sesuai persamaan (1) berikut:

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

dimana $W_{t,d}$ adalah bobot TF-IDF, $tf_{t,d}$ adalah jumlah frekuensi kata, idf_t adalah jumlah *inverse* frekuensi dokumen tiap kata, df_t adalah jumlah frekuensi dokumen tiap kata dan N adalah jumlah total dokumen.

2.3 *K-Fold Cross Validation*

Metode *K-fold cross validation* digunakan untuk membagi data menjadi data latih dan data uji, dengan data uji dapat diambil dari setiap data (Gokgoz and Subasi, 2015). Salah satu alasan metode ini populer digunakan adalah kemampuan metode ini dalam mengurangi bias yang dapat terjadi dalam proses pengambilan sampel angka partisi data. Angka partisi ini digunakan untuk partisi data latih dan data uji.

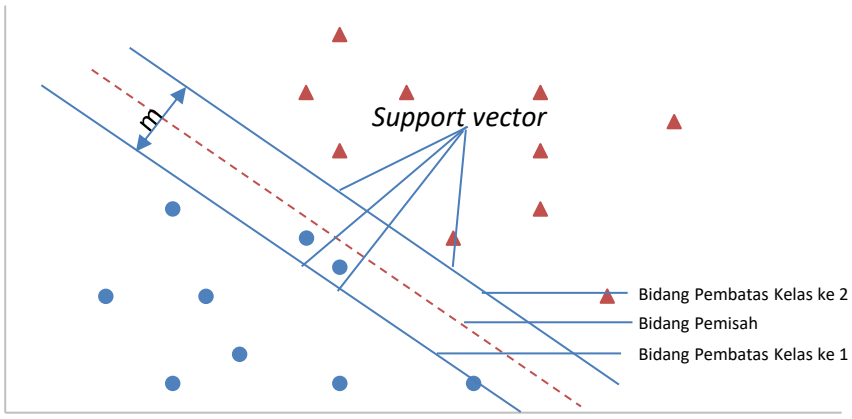
2.4 Mesin Vektor Pendukung Untuk *Linearly Separable Data*

Mesin vektor pendukung merupakan salah satu metode yang dapat digunakan untuk mempelajari area yang memisahkan kategori dalam suatu observasi (Williams, 2011). Salah satu penerapan metode mesin vektor pendukung adalah pemisahan data secara linear, yang dikenal dengan *linearly separable data*. Misalkan terdapat titik data x_i dan y_i yang berperan sebagai label kategori dalam *dataset* (Pisner and Schnyer, 2020).

Gambar 1 menunjukkan sepasang bidang pembatas yang sejajar dengan kelas data yang dipisahkan. *Support vector* adalah data pada bidang pembatas. Persamaan (2) adalah *hyperplane* dapat dituliskan sebagai berikut:

$$f(x) = (\boldsymbol{w^T x}) + b \quad (2)$$

dimana $\boldsymbol{w}$ adalah vektor normal untuk menentukan orientasi dari *hyperplane*. b adalah suatu konstanta skalar sebagai penentu posisi fungsi *hyperplane* terhadap titik asal (Sari, 2017). Dalam kasus ini, 2 *hyperplane* yang sejajar akan dipisahkan agar dapat dibentuk ketidaksamaan model matematika untuk kelas positif dan negatif.



Gambar 1. Sepasang bidang pembatas yang sejajar dengan kelas data yang dipisahkan

Langkah membentuk fungsi *hyperplane* yang optimal pada kasus *linear separable* seperti pada persamaan (3) berikut ini:

$$f(x_d) = \sum_{i=1}^{n} \alpha_i y_i (x_i x_d) + b, \quad (3)$$

dengan x_d adalah data yang akan diklasifikasikan, α_i adalah solusi optimal dari masalah optimasi.

2.5 Metode Kernel pada Mesin Vektor Pendukung *Non Linearly Separable* Data

Peubah *slack* (ξ_i) ditambahkan ke mesin vektor pendukung *nonlinear* dengan $\xi_i \geq 0, i = 1,2,\dots,n$, untuk memperoleh modifikasi *hyperplane* kelas positif dan negatif. Selanjutnya fungsi *hyperplane* optimal untuk kasus mesin vektor pendukung *non-linear* hampir sama dengan kasus *linear separable*. Oleh karena itu, kelas berdasarkan *hyperplane* yang optimal untuk kasus dataset *non-linear separable* dapat dibentuk. *K* adalah salah satu fungsi kernel yang digunakan untuk mengubah data nonlinier menjadi linier (Sari, 2017). Fungsi kernel yang digunakan dalam penelitian ini adalah *Radial Basis Function* (RBF) seperti berikut ini.

$$\mathrm{K}(x_i, x_j) = \exp\left(-\frac{(x_i - x_j)^T (x_i - x_j)}{2\,\gamma^2}\right)$$

2.6 Fitur Berbasis Leksikon

Kata-kata sentimen yang telah diketahui dapat disimpan sebagai kumpulan yang disebut Leksikon. Fitur berbasis leksikon dapat diartikan sebagai fitur atau kata berbobot yang didasarkan kamus atau *lexical*. Fitur berbasis leksikon ini dapat digunakan untuk menentukan arah sentimen sebuah kata (Kurniawan and Indriati, 2019). *Sentiment dictionaries* digunakan untuk pembobotan fitur. Untuk menyeimbangkan bobot fitur berbasis leksikon dengan menggunakan bobot TF-IDF, dilakukan normalisasi fitur jumlah kata yang negatif dan positif menggunakan metode normalisasi *Min-max* (Buntoro *et al.*, 2014; Wahid and Azhari, 2016).

2.7 *Confusion Matrix*

Prediksi kelas suatu data dilakukan dengan menganalisis keakuratan metode klasifikasi. *Confusion matrix* adalah tabel yang dapat digunakan untuk keperluan klasifikasi (Nguyen, *et al*., 2009). Penelitian ini menggunakan klasifikasi biner; positif dan negatif serta *confusion matrix.*

Ukuran yang umum digunakan untuk menghitung ketepatan klasifikasi adalah akurasi, *specificity*, dan *sensitivity/ recall*. Ketepatan pengklasifikasian sebuah dokumen dari data yang seimbang untuk setiap kategori dapat diketahui melalui ukuran akurasi. Selain akurasi, penting juga diketahui keserasian (*Specificity*/*precision*) antar kelas data sebenarnya dengan kelas data prediksi. Ukuran kinerja model juga dapat dilihat berdasarkan keakuratan banyaknya data yang dihasilkan oleh sistem dari kelas yang sebenarnya. Ukuran ini disebut dengan *Sensitivity*. Alternatif untuk mengevaluasi model yang kelas datanya seimbang juga dapat menggunakan gabungan *specificity* dan *sensitivity* yang dikenal dengan nama *F-measure*. Untuk data tidak seimbang pengukuran dapat dilakukan dengan mempertimbangkan nilai *G-mean* dan *Area Under Curve* (AUC) (Bekkar *et al.*, 2013) :

3. HASIL DAN PEMBAHASAN

3.1 Pra-proses Teks

Tahapan awal pada pra-proses teks ini adalah *case folding*, berupa huruf kapital diubah menjadi huruf kecil. Kemudian, dilakukan *data cleansing* yang terdiri dari 9 tahapan, yaitu menghapus *hashtag, space,* digit, dan tanda baca, serta mengubah *username* menjadi [*username*], *emoticon* neg menjadi [neg], *emoticon* positif menjadi [pos], kata tidak baku menjadi baku, dan negasi menjadi [not]. Setelah *data cleansing*, tahapan berikutnya adalah proses *stemming*, *stopword* dan *tokenizing*. Proses *stemming* dilakukan untuk mendapatkan kata dasar. Hasil proses *stemming*, *stopword* dan *tokenizing* masing-masing disajikan pada Tabel 1, 2 dan 3.

Tabel 1 Teks sebelum dan setelah dilakukan *Stemming*

Teks sebelum dilakukan *stemming*	Teks setelah dilakukan *stemming*
ibu dan ibu mertua sudah mendapatkan dosis pertama vaksin covid sedikit rasa lega di tengah pandemi ini [pos]	ibu dan ibu mertua sudah dapat dosis pertama vaksin covid sedikit rasa lega di tengah pandemi ini [pos]

Tabel 2 Teks sebelum dan setelah dilakukan *stopword*

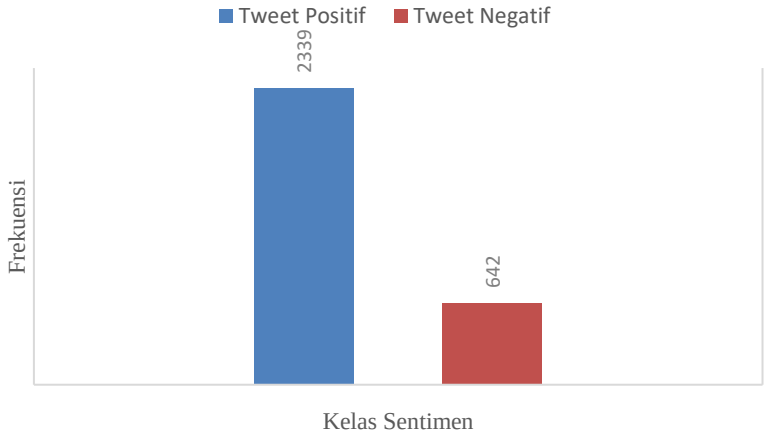
Teks sebelum dilakukan *stopword*	Teks setelah dilakukan *stopword*
ayo vaksin lindungi diri dan keluarga dari covid [username]	ayo vaksin lindungi keluarga covid [username]

Tabel 3 Teks sebelum dan setelah dilakukan *Tokenizing*

Teks sebelum dilakukan *tokenizing*	Teks setelah dilakukan *tokenizing*
ayo vaksin lindungi keluarga covid [username]	[ayo, vaksin, lindungi, keluarga, covid, [username]]

3.2. Deskripsi Data

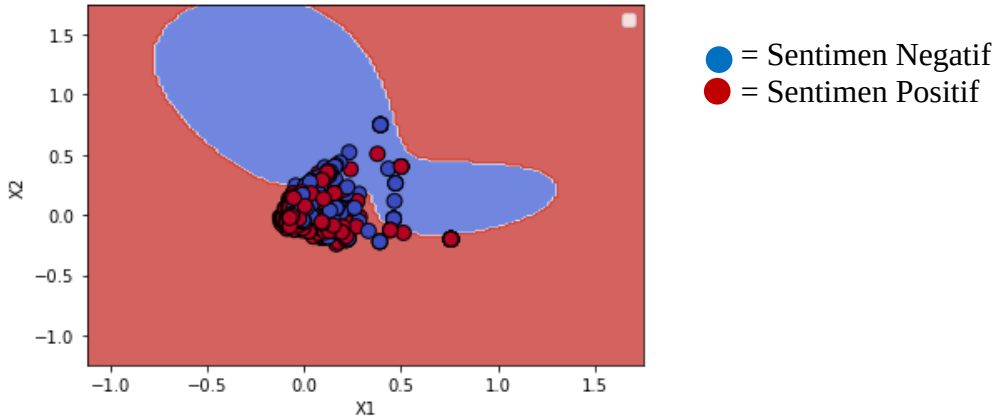
Dari hasil *crawling* pada *twitter* menggunakan *snscrape* pada tanggal 01 Januari sampai 09 April 2021 dengan kata kunci "Vaksinasi COVID-19" diperoleh sebanyak 10.000 data *tweet* menggunakan Bahasa Indonesia. Selanjutnya dilakukan pelabelan secara manual dan diperoleh total data yang telah diberi label yaitu 3046 data dan 6954 data diabaikan karena merupakan data *noise*. Namun setelah dilakukan pengecekan pada program didapat data yang berulang sehingga total data yang diperoleh menjadi 2981 dengan sentimen positif sebanyak 2339 data *tweet* dan sentimen negatif sebanyak 642 data *tweet*. Gambar 2 merupakan diagram batang dari klasifikasi komentar masyarakat pada kebijakan vaksinasi COVID-19 di Indonesia. Pada Gambar 3 itu terlihat adanya ketidakseimbangan kelas sentimen positif dan negative. Oleh karena itu, evaluasi model dapat didasarkan pada nilai akurasi, *G-mean* dan AUC.



Gambar 2 Diagram batang kategori data vaksinasi COVID-19 di Indonesia

3.3 *Term Frequency Inverse Document Frequency*

Pada penelitian ini diperoleh 4438 kata dasar yang berasal dari 2981 *tweet*. Pembobotan TF-IDF pada kata dasar setiap dokumen menyebabkan adanya sebaran data. Dengan menggunakan metode mesin vektor pendukung RBF tanpa fitur leksikon, diperoleh sebaran klasifikasi komentar masyarakat di *twitter* pada kebijakan vaksinasi COVID-19 yang ditunjukkan pada Gambar 3.



Gambar 3 *Scatter Plot* klasifikasi komentar masyarakat di twitter pada kebijakan vaksinasi COVID-19 dengan mesin vektor pendukung kernel RBF tanpa fitur leksikon

Gambar 3 menunjukkan distribusi klasifikasi komentar masyarakat di *twitter* terhadap kebijakan vaksinasi COVID-19 yang tidak dapat dipisahkan secara linier. Oleh karena itu, digunakan pengklasifikasian *non-linear*. Pada Gambar 3 dapat dilihat area data yang bersentimen positif berwarna merah, dan data bersentimen negatif berwarna biru. Pada studi ini digunakan kernel RBF karena persebaran data tidak mendekati bentuk *polynomial* dan *sigmoid* sehingga digunakan RBF.

3.4 Mesin Vektor Pendukung

Pembobotan TF-IDF pada kata dasar setiap dokumen menyebabkan adanya persebaran data yang tidak terpisah secara linear sehingga digunakan klasifikasi *non-linear* dengan kernel RBF. Dalam penelitian ini, data latih sebanyak 80% yang digunakan untuk membentuk model klasifikasi, sedangkan data uji sebanyak 20% digunakan untuk membuat prediksi berdasarkan model klasifikasi yang terbentuk sehingga dapat dievaluasi nilai kebaikan klasifikasinya. Pada tahap pembentukan model digunakan data latih dengan metode mesin vektor pendukung kernel RBF tanpa fitur leksikon. Pada studi ini, nilai parameter C dan gamma yang digunakan masing-masing 10 dan 1 dengan nilai akurasi yang diperoleh sebesar 80%. Nilai gamma sebesar 1 disubtitusikan ke dalam persamaan kernel RBF, sehingga terbentuk fungsi kernel RBF yaitu,

$$K(x_i, x_j) = \exp\left(-0.5 \text{ x } (x_\text{i} - x_\text{j})^T (x_i - x_\text{j})\right)$$

Fungsi kernel yang terbentuk ini sejalan dengan bentuk yang diperoleh oleh Gopi *et al.*, (2020). Selanjutnya, pembentukan fungsi *hyperplane* dilakukan dengan menggunakan fungsi kernel. Fungsi *hyperplane* ini dibentuk dengan mengganti nilai positif dari kategori *support vector* pada x_i dan nilai negatif dari *support vector* pada x_j. Kemudian, dengan mensubstitusikan fungsi kernel diperoleh kategori prediksi untuk setiap *tweet* dengan model sebagai berikut:

$$f(x_d) = \sum_{i=1}^{2981} \alpha_i y_i K(x_i, x_d) + 0{,}54313195$$

dengan α_i adalah nilai *Lagrange Multiplier* dari fungsi klasifikasi kernel nonlinear, y_i adalah kelas data, x_d adalah data yang akan di klasifikasi dan x_i adalah data yang menjadi *support vector* dengan $i = 1,2,\ldots,2981$.

3.5 Fitur Berbasis Leksikon

Dalam penelitian ini kamus sentimen diperoleh dari penelitian (Wahid and Azhari, 2016) dan kamus dari opinion leksikon (Liu, 2012) yang oleh Devid Haryalesmana dimodifikasi dan di*translate* menjadi bahasa Indonesia. Kamus sentimen memiliki skala nilai -5 hingga 5. Hasil dari praproses data selanjutnya akan diberikan bobot untuk setiap kata dan dijumlahkan untuk tiap-tiap dokumen. Pada Tabel 4 disajikan perolehan nilai bobot untuk setiap dokumen.

Tabel 4 Nilai Bobot dari Setiap Dokumen

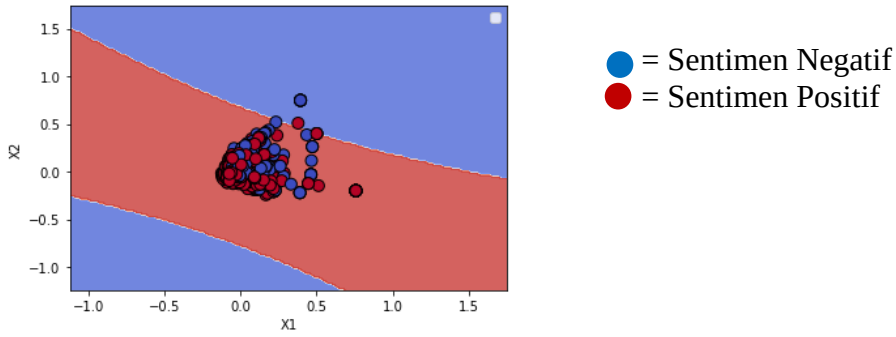
No	*Tweet* sebelum praproses	*Tweet* setelah praproses	*Score*
1	Ayo ikuti anjuran pemerintah untuk VAKSIN. Senagai ikhtiar kita agar terhindar dari Covid 19. Serta tetap jaga protokol	[ayo, ikut, anjuran, pemerintah, untuk, vaksin, ikhtiar, agar, hinder, covid, serta, tetap, jaga, protocol,	2

	kesehatan. Masker. Masker. Masker.	kesehatan, masker, masker, masker]

Setelah diperoleh bobot nilai dari setiap dokumen seperti pada Tabel 4 maka selanjutnya akan dilakukan transformasi menggunakan metode normalisasi *min-max* pada masing-masing dokumen dengan nilai batas bawah dan batas atas masing-masing 0 dan 1. Nilai X_{max} yang diperoleh dalam penelitian ini adalah 16, sedang untuk X_{min} adalah -16. Berikut proses transformasi untuk normalisasi *min-max*,

$$X_1' = \frac{2-(-16)}{16-(-16)}(1-0)+0 = 0{,}562$$

Dari hasil normalisasi selanjutnya data diklasifikasikan dengan ketentuan $X_i < 0.5$ bersentimen negatif dan $X_i \geq 0.5$ bersentimen positif. Dari hasil fitur berbasis leksikon diperoleh data bersentimen positif sebanyak 2101 dan data bersentimen negatif sebanyak 880 berdasarkan bobot nilai dari masing-masing kata. Setelah selesai tahap praproses dan pelabelan data, dilanjutkan ke tahap analisis klasifikasi menggunakan mesin vektor pendukung. Gambar 4 memperlihatkan sebaran klasifikasi komentar masyarakat di *twitter* pada kebijakan vaksinasi COVID-19 menggunakan mesin vektor pendukung dengan fitur berbasis leksikon.



Gambar 4 *Scatter plot* data twitter opini pada vaksinasi COVID-19 di Indonesia dengan mesin vektor pendukung kernel RBF dengan fitur berbasis leksikon

Pelatihan model menggunakan data latih dengan mesin vektor pendukung kernel RBF dengan fitur berbasis leksikon menggunakan nilai parameter untuk C yaitu 1000 dan gamma yaitu 0,01. Hasil akurasi yang diperoleh adalah sebesar 0,90 atau 90%. Selanjutnya untuk membentuk fungsi kernel RBF, nilai gamma yang diperoleh disubstitusikan pada persamaan kernel RBF berikut ini:

$$K(x_i, x_j) = \exp\left(-50 \text{ x } (x_\text{i} - x_\text{j})^T (x_i - x_\text{j})\right)$$

Kemudian, dengan mensubstitusikan fungsi kernel sehingga diperoleh kategori prediksi setiap *tweet* dengan model berikut ini:

$$f(x_d) = \sum_{i=1}^{2981} \alpha_i y_i K(x_i, x_d) + 0{,}20032726$$

3.5 Perbandingan Mesin Vektor Pendukung Dengan dan Tanpa Fitur Leksikon

Untuk mengevaluasi kinerja klasifikasi metode mesin vektor pendukung dengan dan tanpa fitur berbasis leksikon, digunakan *confusion matrix*.

Tabel 5 Evaluasi metode mesin vektor pendukung dengan dan tanpa fitur leksikon untuk klasifikasi komentar masyarakat pada vaksinasi COVID-19 di *twitter*

Metode	Akurasi	*G-Mean*	AUC
Mesin Vektor Pendukung	83%	50%	61,35%
Fitur berbasis leksikon	90%	86,63%	87%

Tabel 5 menunjukkan perbandingan kinerja metode tersebut pada klasifikasi komentar masyarakat tentang vaksinasi COVID-19. Hasil pada Tabel 5 menunjukkan bahwa metode mesin vektor pendukung dengan dengan kernel RBF berbasis fitur leksikon lebih baik daripada tanpa fitur dalam mengukur ketepatan klasifikasi terkait komentar masyarakat tentang vaksinasi COVID-19 di Indonesia.

4. KESIMPULAN

Pada studi ini telah dilakukan analisis komentar masyarakat di *twitter* terhadap kebijakan vaksinasi COVID-19 di Indonesia. Metode mesin vektor pendukung dengan kernel RBF untuk fitur berbasis leksikon dan tanpa fitur leksikon telah digunakan. Kedua metode ini mampu mengklasifikasikan komentar masyarakat pada kebijakan vaksinasi COVID-19 kedalam dua kelas sentimen yaitu sentimen positif dan negatif. Selain itu, evaluasi metode mesin vektor pendukung tanpa fitur leksikon dengan kernel RBF memberikan nilai akurasi, *G-mean*, dan AUC masing-masing sebesar 83%, 50% dan 61,35% pada klasifikasi komentar masyarakat terhadap vaksinasi COVID-19. Selanjutnya, metode mesin vektor pendukung dengan kernel RBF dengan fitur leksikon memberikan nilai akurasi (90%), *G-mean* (86.63%) dan AUC (87%) lebih baik daripada tanpa fitur dalam mengklasifikasi komentar masyarakat pada kebijakan vaksinasi COVID-19 di *twitter*.

DAFTAR PUSTAKA

Bekkar, M., Djemaa, H.K. and Alitouche, T.A. (2013), "Evaluation measures for models assessment over imbalanced data sets", *J Inf Eng Appl*, Vol. 3 No. 10.

Buntoro, G.A., Adji, T.B. and Purnamasari, A.E. (2014), "Sentiment Analysis Twitter dengan Kombinasi Lexicon Based dan Double Propagation", *The 6th Conference on Information Technology and Electrical Engineering (CITEE)*, pp. 39–43.

Dhawan, V. and Zanini, N. (2014), "Big data and social media analytics", *Research Matters: A Cambridge Assessment Publication*, Vol. 18 No. 7, pp. 36–41.

Gokgoz, E. and Subasi, A. (2015), "Comparison of decision tree algorithms for EMG signal classification using DWT", *Biomedical Signal Processing and Control*, Elsevier, Vol. 18, pp. 138–144.

Gopi, A.P., Jyothi, R., Narayana, V.L. and Sandeep, K.S. (2020), "Classification of tweets data based on polarity using improved RBF kernel of SVM", *International Journal of Information Technology*, Springer, pp. 1–16.

Kurniawan, A. and Indriati, S.A. (2019), "Analisis Sentimen Opini Film Menggunakan Metode Naïve Bayes dan Lexicon Based Features", *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer E-ISSN*, Vol. 2548, p. 964X.

Kurniawan, T. (2017), "Implementasi Text Mining Pada Analisis Sentimen Pengguna Twitter terhadap Media Mainstream Menggunakan Naïve Bayes Classifier dan Support Vector Machine", Institut Teknologi Sepuluh Nopember.

Liu, B. (2012), "Sentiment analysis and opinion mining", *Synthesis Lectures on Human Language Technologies*, Morgan & Claypool Publishers, Vol. 5 No. 1, pp. 1–167.

Nugroho, A.S., Witarto, A.B. and Handoko, D. (2003), "Support vector machine",

Proceeding Indones. Sci. Meeting Cent. Japan.
Nguyen, G. H., Abdesselam, B. & Son, P. L. (2009). Learning pattern classification tasks with imbalanced data sets. *Pattern recognition*, pp. 193-208.
Pisner, D.A. and Schnyer, D.M. (2020), "Support vector machine", *Machine Learning*, Elsevier, pp. 101–121.
Sari, P.D. (2017), "Analisis Credit Scoring Menggunakan Regresi Logistik LASSO dan Support Vector Machine (SVM).", Bogor Agricultural University (IPB).
Septian, J.A., Fachrudin, T.M. and Nugroho, A. (2019), "Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor", *INSYST: Journal of Intelligent System and Computation*, Vol. 1 No. 1, pp. 43–49.
Vidya, N.A., Fanany, M.I. and Budi, I. (2015), "Twitter sentiment to analyze net brand reputation of mobile phone providers", *Procedia Computer Science*, Elsevier, Vol. 72, pp. 519–526.
Wahid, D.H. and Azhari, S.N. (2016), "Peringkasan sentimen ekstraktif di twitter menggunakan hybrid TF-IDF dan cosine similarity", *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, Vol. 10 No. 2, pp. 207–218.
Williams, G. (2011), *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery*, Springer Science & Business Media.