Jurnal Ilmu Komputer & Agri-Informatika
Volume 9 Nomor 1 halaman 79 – 89
e ISSN: 2654-9735, p ISSN: 2089-6026

Tersedia secara *online* di:
http://journal.ipb.ac.id/index.php/jika

Analisis Sentimen Pengguna Twitter Terhadap Program Vaksinasi Covid-19 di Indonesia Menggunakan Algoritme Support Vector Machine

Sentiment Analysis of Twitter Users on COVID-19 Vaccination Program in Indonesia using Support Vector Machine Algorithm

QARRY ATUL CHAIRUNNISA[1], YENI HERDIYENI[1], MEDRIA KUSUMA DEWI HARDHIENATA[1*], JULIO ADISANTOSO[1]

Abstrak

Kebijakan vaksinasi COVID-19 di Indonesia menimbulkan pro dan kontra. Pemerintah harus mengevaluasi alasan masyarakat yang kontra terhadap kebijakan tersebut, agar program vaksinasi dapat berjalan dengan lancar. Analisis sentimen sebagai cara untuk melihat polaritas opini, memungkinkan untuk mengklasifikasi tanggapan positif, negatif maupun netral di Twitter terkait kebijakan vaksinasi tersebut. Penelitian ini bertujuan untuk mengetahui tanggapan masyarakat terhadap vaksinasi COVID-19 dengan melihat distribusi kata dan membuat model klasifikasi *support vector machine* (SVM). Analisis sentimen terdiri dari beberapa tahapan yaitu pengumpulan data, praproses data, pembobotan data, analisis data, pembagian data, pemodelan klasifikasi, *hyperparameter tuning*, dan evaluasi model. Model yang dihasilkan menunjukkan kinerja yang cukup optimal dalam mengklasifikasi sentimen dengan akurasi, presisi, *recall*, dan *f1-score* sebesar 90%. Hasil dari sentimen analisis yang diperoleh ialah berupa gagasan, keluhan, dan saran terhadap program vaksinasi COVID-19.

Kata kunci**:** analisis sentimen, COVID-19, *support vector machine*, vaksinasi, twitter

Abstract

The COVID-19 vaccination policy in Indonesia turns out to be both pros and cons. The government must evaluate the underlying reason of why some people are against the policy, so that the vaccination program can run smoothly. Sentiment analysis as a way to see the polarity of opinion, makes it possible to classify positive, negative or neutral responses on Twitter regarding the vaccination policy. This study aims to determine the public's response to COVID-19 vaccination by examining word distribution and creating a support vector machine (SVM) classification model. The sentiment analysis conducted in this study consists of several stages, namely data collection, data preprocessing, data weighting, data analysis, data sharing, classification modeling, hyperparameter tuning and model evaluation. The results of this study are a model with a relatively optimal performance in classifying sentiment with an accuracy, precision, recall and f1-score of 90%. The results of the sentiment analysis obtained are in the form of ideas, complaints, and suggestions regarding the COVID-19 vaccination program.

Keywords: COVID-19, sentiment analysis, support vector machine, vaccination

PENDAHULUAN

Penyakit virus Corona baru atau yang disebut dengan SARS-CoV-2 telah dilaporkan pertama kali ditemukan di China pada bulan Desember 2019 (Xueting *et al.* 2020). Pada tanggal 2 Maret 2020, kasus pertama virus tersebut ditemukan di Indonesia (Nuraini 2020). Virus Corona terus menyebar, hingga pada tanggal 19 Oktober 2020 Indonesia ditetapkan

[1]Departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor 16680
*Penulis Korespondensi. Surel: medria.hardhienata@apps.ipb.ac.id

sebagai negara nomor satu dengan tingkat kematian tertinggi se-Asia Tenggara (Shalihah 2020). Sejak menyebarnya virus Corona di Indonesia, masyarakat miskin, rentan miskin dan yang bekerja di sektor informal menjadi aspek masyarakat yang paling terdampak disebabkan oleh adanya pemutusan hubungan kerja serta penurunan upah gaji (BPS 2020). Berdasarkan data tersebut, pemerintah harus mengambil tindakan yang cepat dan tepat dalam menangani pandemi ini terutama pada sektor perekonomian.

Salah satu kebijakan pemerintah agar tercapainya kekebalan komunitas sehingga sektor perekonomian bisa kembali stabil adalah pemberian vaksinasi Sinovac gratis untuk seluruh masyarakat berdasarkan urutan prioritas (Dewi 2020). Meskipun demikian, kebijakan vaksinasi gratis ini ternyata menimbulkan pro dan kontra di kalangan masyarakat. Banyak masyarakat yang mempertanyakan keampuhan vaksin tersebut. Hal ini ditandai dengan *hashtag* #TolakDiVaksinSinovac yang menjadi *trending* Twitter pada tanggal 12 Januari 2021 (Hidayatullah 2021).

Media sosial kini telah menjadi platform publik yang kuat dalam berbagi pendapat dengan banyak orang, terutama Twitter. Studi yang ada menunjukkan bahwa data Twitter dapat memberikan informasi yang berguna untuk penyakit epidemik, termasuk melacak sentimen, mengukur minat, dan kekhawatiran publik serta melacak tingkat penyebaran penyakit yang dilaporkan (Jia *et al.* 2020). Pada umumnya, pemerintah melakukan jajak pendapat untuk melihat evaluasi dari kebijakan yang ada melalui survei secara langsung dari rumah ke rumah atau dengan pendekatan *person to person* (Gong *et al.* 2020). Di Indonesia sendiri hal ini biasanya dilakukan oleh Badan Pusat Statistika (BPS). Namun, survei ini memiliki banyak kekurangan. Beberapa di antaranya ialah terkait biaya, efektivitas waktu, *human error* dan kurang praktisnya jika ada perubahan-perubahan parameter (Gong *et al.* 2020). Maka dari itu, adanya sosial media Twitter dan hadirnya *machine learning* diharapkan mampu mempermudah dan mengoptimasi pengambilan hasil jajak pendapat yang dibutuhkan.

Penelitian sebelumnya oleh Yulita *et al.* (2021) membahas analisis sentimen terhadap vaksinasi COVID-19 pada pengguna Twitter menggunakan algoritme klasifikasi Naive Bayes. Penelitian tersebut memberikan kesimpulan bahwa masyarakat yang menentang vaksinasi lebih sedikit daripada masyarakat yang setuju dan mau mengikuti vaksinasi. Dalam penelitian Yulita *et al.* (2021) digunakan data sebesar 3780 *tweet* dan algoritme yang digunakan bekerja optimal untuk data dengan jumlah sedikit. Penelitian oleh Rahman dan Permana (2020) mengenai analisis sentimen Twitter terkait vaksinasi COVID-19 juga telah dilakukan dengan kata kunci "vaksin covid" dan "vaksin corona". Pada penelitian tersebut digunakan metode *lexicon-based* untuk memisahkan sentimen pro dan kontra dengan pengelompokan topik pembicaraan masyarakat menggunakan *latent disrichlet allocation* (LDA). Penelitian sebelumnya mengenai vaksinasi COVID-19 dengan kombinasi kata kunci "vaccine" dengan kata kunci lainnya seperti "covid", "coronavirus", "ncov2019", dan "SARS-CoV-2" telah dilakukan dengan melakukan pengelompokan sentimen agar dapat mengetahui misinformasi mengenai vaksin yang beredar di masyarakat (Yousefinaghani *et al.* 2021).

Berbeda dengan penelitian sebelumnya yang telah disebutkan di atas, penelitian ini bertujuan untuk melakukan analisis sentimen positif, negatif, netral masyarakat pengguna Twitter terhadap tren kebijakan vaksinasi COVID-19 dengan menggunakan algoritme klasifikasi *support vector machine* (SVM) dengan korpus berbahasa Indonesia. Analisis sentimen pengguna Twitter terhadap vaksinasi COVID-19 di Indonesia dilakukan dalam penelitian ini untuk melihat polaritas opini masyarakat terhadap kebijakan vaksinasi yang diluncurkan oleh pemerintah. Hal ini penting dilakukan agar opini masyarakat terkait kebijakan vaksin dapat dianalisis dengan cepat sebagai bahan pertimbangan bagi pengambil kebijakan kedepannya.

Metode SVM dalam penelitian ini dipilih karena metode ini telah terbukti dalam penelitian sebelumnya dapat memberikan hasil akurasi yang baik dalam klasifikasi, contohnya pada penelitian Shivaprasad dan Shetty (2017). Berdasarkan penelitian tersebut, metode SVM terbukti dapat memberikan akurasi yang baik dalam melakukan analisis sentimen dengan hasil

akurasi sebesar 98% dalam tinjauan sebuah produk. Metode SVM juga telah menunjukan hasil baik dalam sentimen analisis dalam penelitian Pang *et al*. (2002). Selain itu, pendekatan SVM juga telah terbukti memiliki keefektifan tinggi dalam kategori teks serta mampu mengungguli pendekatan *machine learning* lain seperti pendekatan Naive Bayes (Joachims 1998). SVM memiliki kelebihan dalam menentukan jarak pemisah antar kelas dengan menggunakan *support vector* sehingga proses komputasi yang dilakukan pada data yang cukup besar dapat menjadi lebih cepat (Vapnik dan Cortes 1995). Dengan perbedaan jumlah data yang jauh lebih besar untuk memuat lebih banyak tanggapan masyarakat luas, penelitian ini akan mengklasifikasi sentimen menggunakan algoritme SVM.

METODE

Penelitian ini dilakukan melalui beberapa tahap, yaitu pengambilan data, klasifikasi awal, praproses data, analisis sentimen, pembobotan kata, pembagian data, pemodelan klasifikasi, *hyperparameter tuning*, dan evaluasi. Tahap penelitian tersebut dapat dilihat pada Gambar 1.

Penelitian ini akan mengklasifikasikan data menjadi tiga kelas yaitu sentimen positif, negatif dan netral. Data *tweet* yang digunakan adalah tweet dengan kata kunci "vaksin corona" yang hanya mengambil fitur isi *tweet*, *retweetCount*, *likeCount*, dan tanggal. Waktu *tweet* yang diambil ialah sejak 8 September 2020 sampai 1 Juni 2021.

Pengambilan Data

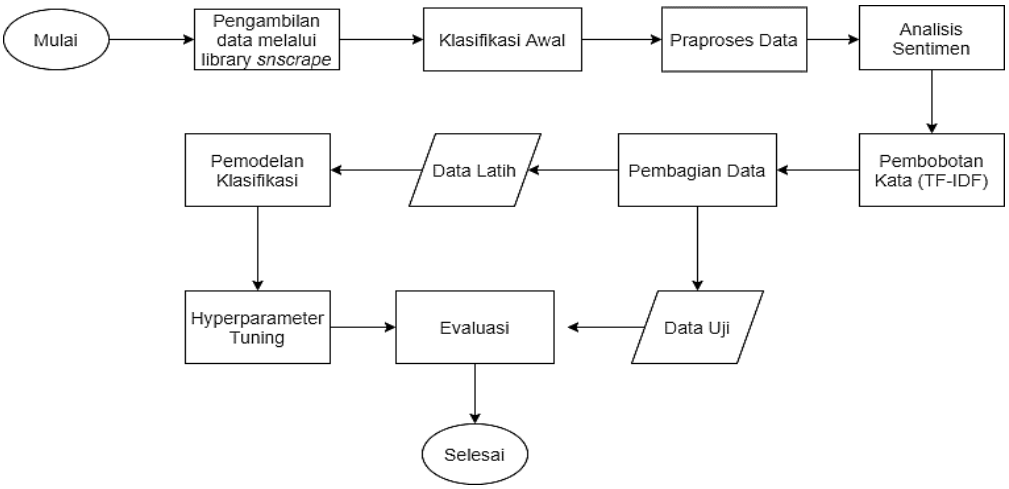
Tahap pertama yang dilakukan adalah mengumpulkan *tweet* dengan kata kunci "vaksin corona". Data diekstrak menggunakan *Package snscrape*.

Klasifikasi Awal

Sebelum dilakukan praproses, data *tweet* akan diklasifikasikan secara manual berdasarkan label positif, negatif dan netral. Pelabelan dilakukan secara manual agar lebih akurat dalam memahami sentimen secara utuh dengan melihat emotikon yang ada dan kata-kata yang disingkat

Praproses Data

Praproses perlu dilakukan agar proses klasifikasi lebih cepat dan *vektor space model* menjadi lebih rendah. Tujuan dilakukannya praproses data *tweet* adalah untuk menyamakan standar bentuk kata, menyedikitkan kecacatan data, serta memudahkan pencocokan kata karena adanya pengurangan volume kata. Tahapan yang dilakukan pada praproses ada beberapa langkah yaitu *case folding*, *data cleaning, stopword removal*, tokenisasi, normalisasi, dan *stemming.*



Gambar 1 Tahapan penelitian.

Pembobotan Kata

Pembobotan kata dilakukan dengan menggunakan algoritme TF-IDF dengan menghitung jumlah frekuensi kata pada dokumen (Salton dan Buckley 1998; Berger *et al.* 2000). Persamaan untuk menghitung TF-IDF adalah sebagai berikut:

$$w_{i,j} = tf_{i,j} \times log_2(\frac{N}{df_i}), \quad (1)$$

dengan $tf_{i,j}$ adalah jumlah term i dalam dokumen j, df_i adalah jumlah dokumen yang mengandung i, sedangkan N adalah jumlah keseluruhan dokumen.

Analisis Sentimen

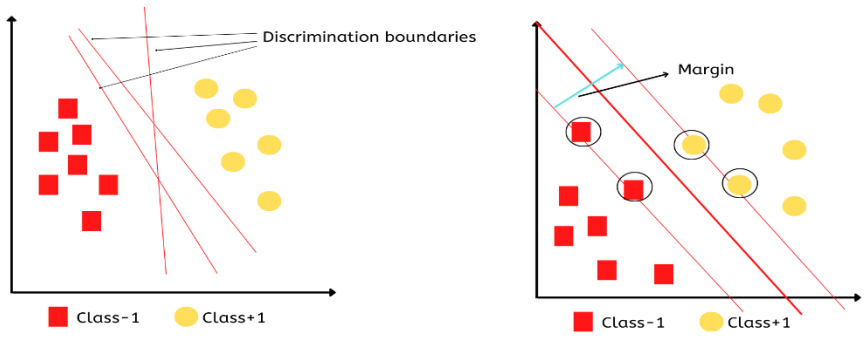
Dalam tahap analisis sentimen, dokumen akan terbagi menjadi kelompok kata positif, negatif, netral dan akan menunjukkan reaksi masyarakat pengguna *twitter* terhadap program vaksinasi Covid-19. Pada penelitian terdahulu, analisis sentimen juga sering digunakan sebagai alat ukur reaksi masyarakat terhadap turbulensi politik. Beberapa di antaranya ialah seperti penelitian prediksi dan analisis pemilihan presiden Indonesia 2019 (Budiharto dan Meiliana 2018) dan opini publik terhadap COVID-19 di California dan New York (Xueting *et al.* 2020).

Pembagian Data

Pada tahap pembagian data, data akan dibagi menjadi data latih dan data uji menggunakan 4-*fold cross validation.* Pemilihan 4-*fold cross validation* dalam kasus ini dipilih dikarenakan nilai ini telah diuji menghasilkan akurasi yang cukup baik.

Pemodelan Klasifikasi

Pemodelan klasifikasi *tweet* akan dilakukan menggunakan algoritme SVM. Pada tahun 1990-an, Vapnik mengusulkan algoritme klasifikasi pembelajaran terawasi (*supervised learning*) yang efisien untuk mesin vektor pendukung, kemudian diterapkan di banyak bidang dan menjadi salah satu algoritme yang paling berguna dalam klasifikasi data (Diya dan Yixi 2019). Algoritme tersebut adalah *support vector machine* (SVM). Ide dasar pembelajaran SVM adalah menyelesaikan *hyperplane* pemisah yang dapat membagi set data pelatihan dengan benar dan memiliki interval geometris terbesar (Diya dan Yixi 2019). *Hyperplane* berguna dalam memisahkan dua kelompok kelas di mana setiap kelas memiliki *pattern* khas tersendiri (Luqyana *et al.* 2018). Dalam hal ini, SVM akan melakukan proses klasifikasi dengan menggunakan *hyperplane* yang ditemukan pada data. Hyperplane terbaik akan menentukan tingkat akurasi dari proses pengklasifikasian pada analisis sentimen. Ilustrasi *hyperplane* pada SVM ditunjukkan pada Gambar 2.



Gambar 2 SVM bekerja dengan menemukan hyperplane terbaik yang memisahkan kedua class -1 dan +1 (Nugroho *et al*. 2003).

Pendekatan SVM yang digunakan dalam penelitian ini adalah metode SVM *one-versus-one* (OVO) yakni sebuah pendekatan di mana permasalahan dibagi menjadi *Binary problem* yang mengkombinasi seluruh kemungkinan pasangan kelas (Galar *et al.* 2011). Metode SVM

OVO ini dipilih dengan pertimbangan bahwa persebaran data yang digunakan relatif seimbang dan berjumlah cukup besar. Dalam penelitian ini, digunakan tiga kelas yakni, kelas positif, kelas negatif dan kelas netral.

Perhitungan matematis utama dalam metode SVM yaitu menggunakan model linear sebagai *decision boundary* dengan bentuk sebagai berikut (Ben-Hur 2010):

$$y(x) = w^T \Phi(x) + b. \quad (2)$$

Dalam Persamaan 2, *x* adalah vektor input, *w* adalah parameter bobot, $\Phi(x)$ adalah fungsi basis dan *b* adalah bias.

Hyperparameter Tuning

Hyperparameter tuning dilakukan untuk mencari parameter yang dapat memberikan performa paling optimal terhadap model. Beberapa *hyperparameter* yang akan digunakan pada model SVM kali ini adalah sigma (C), gamma dan kernel *radial basis function* (RBF).

Evaluasi Model

Pada penelitian ini, data uji akan digunakan untuk mengevaluasi model klasifikasi yang telah digunakan. Pengujian model akan dilakukan dengan menghitung metrik akurasi, spesifisitas dan sensitivitas berdasarkan *confusion matrix* sebagai berikut (Pedregosa *et al* 2011):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (3)$$

$$Precision = \frac{TP}{TP + FP}, \quad (4)$$

$$Recall = \frac{TP}{TP + FN}, \quad (5)$$

$$F\text{-}Measure = \frac{2 + Precision + Recall}{Precision + Recall}. \quad (6)$$

Pada Persamaan 3-6 *true postive* (TP) merupakan jumlah data positif yang terklasifikasi dengan benar, *true negative* (TN) merupakan jumlah data negatif yang terklasifikasi dengan benar, *false positive* (FP) merupakan jumlah data positif yang salah terklasifikasi, sedangkan *false negative* (FN) merupakan jumlah data negatif yang salah terklasifikasi.

HASIL DAN PEMBAHASAN

Pengumpulan Data

Data *tweet* yang berhasil diperoleh ialah berjumlah 42 162 *tweet* yang terdiri dari 4 atribut. Atribut ini yaitu isi *tweet*, tanggal *tweet*, jumlah *retweet* dan jumlah *like* dari *tweet*.

Klasifikasi Awal

Sebelum memasuki tahap praproses, masing-masing data *tweet* diklasifikasikan terlebih dahulu menjadi 3 label data *tweet*. Keterangan dari masing-masing label dapat dilihat pada Tabel 1.

Tabel 1 Keterangan Pelabelan Data *Tweet*

Label	Keterangan
Positif (1)	Pernyataan setuju berupa ajakan, dukungan, dan informasi pencerdasan ke sesama masyarakat
Negatif (-1)	Pernyataan berupa tolakan, ejekan, dan sindiran
Netral (0)	Pernyataan netral, tidak mengarah kepada dukungan maupun tolakan

Praproses Data

Awalnya, data *tweet* yang digunakan adalah berjumlah 42 163 *tweet* dengan 7 261 label positif, 4 963 label negatif dan 29 939 label netral. Kemudian memasuki tahap *case folding* lalu dilakukan proses pembersihan dengan, menghapus data *tweet* yang duplikat, menghapus emotikon, link *url*, *mention*, *hastag*, gambar dan angka. Setelah proses pembersihan, dilakukan tahap normalisasi, tokenisasi, *stopwords removal* dan *stemming*. Tahapan praproses ini membuat jumlah data *tweet* berkurang menjadi 30 881 *tweet* dengan 5 521 label positif, 4 409 label negatif dan 20 951 label netral. Hasil dari tahap praproses dapat dilihat pada Tabel 2.

Pembobotan Kata

Pada tahap ini, data *tweet* yang telah berbentuk potongan kata diboboti. Gambar 3 menunjukkan perhitungan TF-IDF pada salah satu data.

Analisis Sentimen

Data *tweet* yang berlabel positif dan negatif dikelompokkan masing-masing menjadi 5 *cluster* untuk mengetahui topik sentimen seperti apa yang ramai diperbincangkan masyarakat. Pemilihan 5 *cluster* dalam penelitian ini dilakukan berdasarkan hasil analisis dari Metode Elbow untuk penentuan jumlah *cluster* terbaik (Bholowalia dan Kumar 2014).

***Cluster Tweet* Positif**. Pada *cluster* 1, masyarakat menunjukkan tentang harapan agar pandemi segera berlalu, masyarakat bisa selalu sehat dan vaksin segera hadir agar bisa lebih merasa bebas (aman). Pada *cluster* 2, masyakarat saling mendukung dan menguatkan argumen mengenai vaksin sinovac yang halal serta menyatakan keinginannya untuk vaksin asalkan Presiden Jokowi men-gratiskan. Pada *cluster* 3, masyarakat menyatakan vaksin yang aman untuk lansia dan mengingatkan untuk tetap patuh mengikuti protokol kesehatan meskipun sudah vaksin. Pada *cluster* 4, masyarakat banyak memberitahu bahwa walaupun sudah vaksin masih bisa kena virus corona dan juga menyindir orang-orang yang tidak mau divaksin. Pada *cluster* 5, terdapat opini dari *buzzer* yang agak sulit dibedakan satu dengan lainnya mengenai Polri yang akan mengkawal ketat distribusi vaksin untuk mencegah penularan virus, tujuannya membuat masyarakat merasa aman.

Tabel 2 Contoh tahap praproses data *tweet*

Tweet awal	*Case folding* dan pembersihan	*Stemming*
@AdDien90 Maaf Ustadz🙏, terlepas daripada Vaksin corona itu Konspirasi,Di Negara sayaмcsaya dan banyak lagi rakyat yg tdk percaya dengan Vaksin Corona.karna banyak pula orang sehat yg sdh divaksin mati, bahkan anggota parlemen yg sudah 2 kali vaksin pun skrang jd suspek corona.	maaf ustadz terlepas daripada vaksin corona itu konspirasidi negara sayasaya dan banyak lagi rakyat yg tdk percaya dengan vaksin coronakarna banyak pula orang sehat yg sdh divaksin mati bahkan anggota parlemen yg sudah kali vaksin pun skrang jd suspek corona	['vaksin', 'corona', 'maaf', 'ustadz', 'lepas', 'konspirasi', 'negara', 'saya', 'rakyat', 'tidak', 'percaya', 'corona', 'orang', 'sehat', 'sudah', 'vaksin', 'mati', 'anggota', 'parlemen', 'kali', 'sekarang', 'jadi', 'suspek']

***Cluster Tweet* Negatif**. Pada *cluster* 1**,** masyarakat cenderung menolak vaksin karena banyaknya orang yang mati setelah divaksin. Pada *cluster* 2, masyarakat menolak divaksin karena meragukan efikasinya dengan melihat adanya orang yang sudah divaksin tapi tetap terkena corona. Pada *cluster* 3, banyak yang membicarakan terkait inkonsistensi pernyataan petinggi negara yang awal muncul corona mengeluarkan pernyataan bahwa tidak perlu takut akan corona, namun sekarang meminta rakyat untuk melakukan vaksinasi. Pada *cluster* 4, masyarakat menolak divaksin karena merasa pandemi Covid 19 dan vaksin adalah bisnis dari pejabat dan juga negara china. Masyarakat juga ragu karena inkonsistensi terkait ajakan vaksin namun negara asal vaksin sendiri meragukan efikasinya. Pada *cluster* 5, masyarakat takut

dengan vaksin karena vaksin dipercaya sebagai bisnis, namun akan berusaha lebih yakin jika Pak Jokowi dan jajarannya divaksin terlebih dahulu untuk melihat efektifitasnya.

```
Show TFIDF sample ke-0

['corona', 'okay', 'covid', 'duduk', 'keluarga', 'virus', 'salah', 'contoh', 'merssars', 'serang', 'sistem', 'akut', 'nafas',
'vaksin', 'sudah', 'kaji', 'dengan', 'bantu', 'mrna', 'sedia', 'teknologi', 'maju', 'orang', 'kaya', 'dunia', 'sumbang', 'lag
i']

                        TF              IDF             TF-IDF          Term

array position 89       0.047619        5.900371        0.280970        bantu
array position 175      0.047619        6.631985        0.315809        contoh
array position 176      0.047619        1.119980        0.053332        corona
array position 178      0.047619        2.440845        0.116231        covid
array position 201      0.047619        5.193172        0.247294        dengan
array position 224      0.047619        6.301130        0.300054        duduk
array position 229      0.047619        4.870078        0.231908        dunia
array position 434      0.047619        6.042022        0.287715        kaya
array position 442      0.047619        6.329704        0.301414        keluarga
array position 490      0.047619        6.259745        0.298083        lagi
```

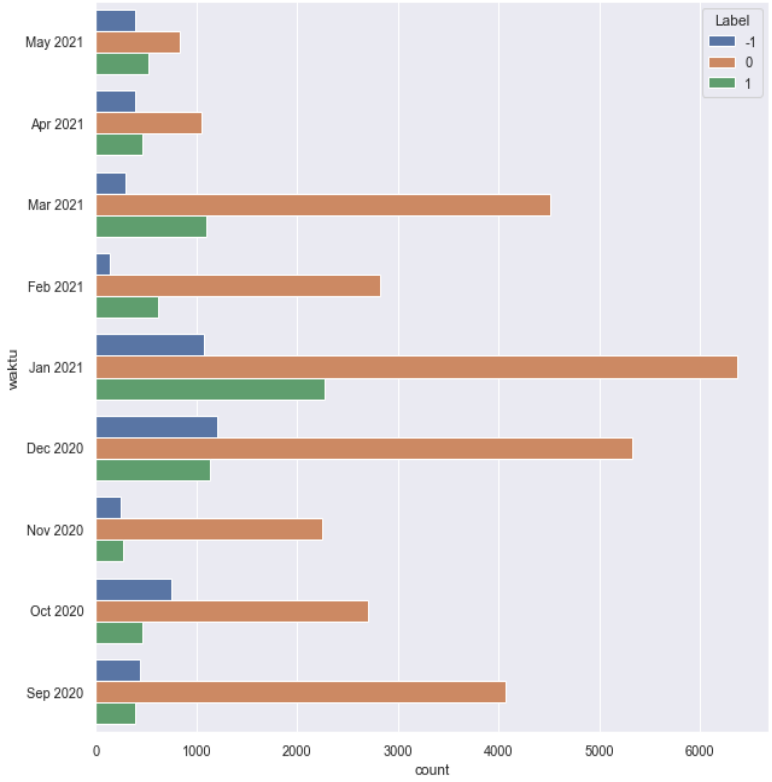
Gambar 3 Contoh perhitungan TF-IDF pada data indeks ke 0.

***Tweet* Netral.** Sebagian besar *tweet* yang bersentimen netral ini berisi cuitan isi berita dari berbagai portal berita seperti Kumparanews yang menginformasikan terkait perkembangan uji klinis vaksin, ditemukannya vaksin, tibanya vaksin, dan sebagainya.

Gambar 4 Dari atas kiri ke kanan, bawah kiri ke kanan: *wordcloud cluster* 1, 2, 3, 4, dan 5.

Gambar 5 Dari atas kiri ke kanan, bawah kiri ke kanan: *wordcloud cluster* 1,2, 3, 4, dan 5.

Gambar 6 *Wordcloud tweet* netral.



Gambar 7 Diagram distribusi *tweet* berdasarkan label dan bulan.

Berdasarkan diagram pada Gambar 7, dapat diketahui bahwa sentimen negatif sangat tinggi di bulan Desember 2020 bertepatan dengan dikeluarkannya kebijakan vaksin pada tanggal 16 Desember 2020. Namun, sentimen negatif menurun di bulan Januari 2021 bertepatan dengan Presiden Jokowi menjadi orang pertama yang divaksin pada tanggal 13 Januari 2021. Setelah dilakukan analisis sentimen, terdapat beberapa saran sebagai berikut:

- Perlu adanya penyampaian informasi secara utuh untuk menjelaskan efektivitas vaksin dan penjaringan hoax tentang vaksin.
- Narasi berita terkait hal-hal yang bisa memicu stigma negatif terhadap vaksin perlu diperbaiki agar masyarakat tidak mudah memberikan kesimpulan.
- Media perlu berhati-hati menyatakan sikap sehingga masyarakat tidak memandangnya sebagai sebuah inkonsistensi.
- Perlu memperbanyak sosialisasi terkait apa yang terjadi di kalangan masyarakat seperti mengapa seseorang yang sudah bisa vaksin masih bisa positif dan penjelasan detail terkait efek vaksin di tubuh manusia. Hal ini bisa dilakukan secara serentak di media-media edukasi.

Pembagian Data

Kurang seimbangnya proporsi data pada tiap label diatasi dengan melakukan uji coba pada beberapa algoritme seperti *random oversampling* (ROS), *random undersampling* (RUS), dan SMOTE (Lema *et al.* 2017). Uji coba pertama dilakukan menggunakan ROS dan menghasilkan jumlah data sebesar 62 853 *tweet* dengan proporsi masing-masing kelas sebesar 20 951 *tweet*. Kemudian uji coba kedua dilakukan dengan mengombinasi ROS-RUS dan menghasilkan jumlah data sebesar 15 451 *tweet* dengan proporsi kelas positif 5151 *tweet* dan kelas negatif serta netral masing-masing berjumlah 5150 *tweet*. Setelah itu dilakukan uji coba ketiga dengan menggunakan algoritme SMOTE yang menjadikan jumlah data *tweet* bertambah menjadi 62 853 *tweet*. Pada akhirnya, penelitian ini akan menggunakan data yang telah diproses menggunakan algoritme SMOTE. Setelah itu, dilakukan pembagian data *tweet* menjadi data latih sebesar 47 410 *tweet* dan data uji sebesar 15 713 *tweet* menggunakan *k-fold cross validation* dengan *k* sebesar 4.

Pemodelan Klasifikasi

Model klasifikasi *tweet* dibangun menggunakan *package scikit-learn* bernama *support vector classification* (SVC).

Hyperparameter Tuning

Hyperparameter tuning dilakukan menggunakan metode *grid search cross validation*. Ruang pencarian untuk tiap *hyperparameter* adalah sigma (C) dengan pilihan nilai 0.1, 1, 10, 100, 1000, dan gamma dengan pilihan nilai 1, 0.1, 0.01, 0.001, 0.0001. *Hyperparameter tuning* ini menghasilkan sigma (C) dengan nilai 1000 dan gamma dengan nilai 1. Kemudian setelah itu dilakukan proses pelatihan menggunakan *hyperparameter* yang dihasilkan.

Evaluasi Model

Setelah dilakukan pemodelan menggunakan *hyperparamater* terbaik, model dievaluasi berdasarkan nilai akurasi. Nilai akurasi yang digunakan dalam penelitian ini adalah pada rentang 0-1. Berikut hasil akurasi dengan beberapa percobaan algoritme penyeimbangan data ROS, ROS-RUS, dan SMOTE yang terdapat pada Tabel 3.

Perhatikan bahwa meskipun Algoritme ROS memiliki nilai akurasi yang lebih tinggi dari algoritme SMOTE (Tabel 3), Algoritme ROS tidak dipilih dalam penelitian ini karena tidak merepresentasikan persebaran data yang seimbang. Oleh karenanya pendekatan SMOTE dipilih dalam penelitian ini yang memiliki akurasi yang relatif baik, yakni sekitar 0.897.

Hasil evaluasi model ditunjukkan pada Tabel 4. Tabel 4 menunjukkan akurasi pada model yang sudah dilatih yaitu sebesar 0.90. Sedangkan kolom *precision* pada tabel menunjukkan keseluruhan data *tweet* yang terklasifikasi kelas negatif dan netral, terdapat masing-masing sebesar 0.91 dari data *tweet* yang benar-benar termasuk kelas negatif dan netral. Begitu pula pada keseluruhan data *tweet* yang terklasifikasi kelas positif, terdapat 0.88 dari data *tweet* yang benar-benar termasuk kelas positif. Kolom *recall* pada tabel menunjukkan keseluruhan data *tweet* kelas negatif terdapat 0.93 data yang terklasifikasi negatif. Pada data *tweet* kelas positif terdapat 0.95 data yang terklasifikasi positif. Kemudian pada *tweet* kelas netral terdapat 0.82 data yang terklasifikasi netral. Kolom *F1-Score* menunjukkan seberapa baik nilai *precision* dan *recall* pada model ini yaitu sebesar 0.92 untuk kelas negatif, 0.86 untuk kelas netral dan 0.91 untuk kelas positif.

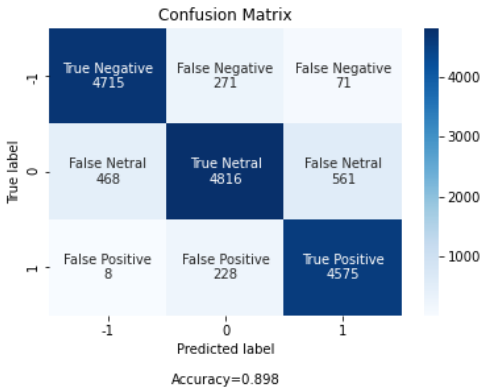
Evaluasi Model dilakukan menggunakan data uji dengan total 15 713 *tweet*. Gambar 8 menunjukkan *confusion matrix* menggunakan pemodelan SVM yang didapatkan dari data latih. Berdasarkan *confusion matrix*, terdapat sebanyak 5057 *tweet* berlabel negatif dengan 4715 terprediksi benar, 271 terprediksi kelas netral, dan 71 terprediksi kelas positif. Selanjutnya sebanyak 5845 *tweet* berlabel netral dengan 4816 terprediksi benar, 468 terprediksi kelas negative, dan 561 terprediksi kelas positif. Sedangkan sebanyak 4811 *tweet* berlabel positif dengan 4575 terprediksi benar, 228 terprediksi netral, dan 8 terprediksi negatif.

Tabel 3 Hasil akurasi berdasarkan beberapa percobaan

Metode penyeimbangan data	Akurasi dalam skala 0-1
ROS	0.943
ROS-RUS	0.666
SMOTE	0.897

Tabel 4 Hasil klasifikasi

	Precision	Recall	F1-Score	Support
-1	0.91	0.93	0.92	5057
0	0.91	0.82	0.86	5845
1	0.88	0.95	0.91	4811
Accuracy			0.90	15713
Macro avg	0.90	0.90	0.90	15713
Weighted avg	0.90	0.90	0.90	15713



Gambar 8 *Confusion matrix* (-1: negatif, 0: netral, 1: positif).

SIMPULAN

Dalam penelitian ini, model SVM berhasil mengklasifikasi sentimen pengguna Twitter terhadap vaksinasi Covid-19 di Indonesia dengan kemampuan mengklasifikasi sentimen yang baik. Hal ini disebabkan oleh jumlah data yang besar dan seimbangnya proporsi setiap kelas sehingga menghasilkan performa model dengan rincian akurasi sebesar 90%, rataan *f1-score* sebesar 90%, rataan *recall* sebesar 90%, dan rataan presisi sebesar 90%.

Hasil analisis sentimen menunjukkan pada bulan Desember 2020 ke Januari 2021, tren *tweet* kelas positif yang berisi pernyataan setuju berupa ajakan dan dukungan terhadap program vaksin COVID-19 meningkat dan tren *tweet* kelas negatif yang berisi pernyataan kontra berupa tolakan dan ejekan terhadap program vaksin COVID-19 menurun. Hal ini juga berkaitan dengan pengumuman pelaksanaan vaksinasi pada bulan Desember 2020 sehingga masyarakat cenderung khawatir, mulai tersebar berbagai isu hingga banyak yang memutuskan enggan untuk divaksin. Selanjutnya pada bulan Januari 2021, *tweet* negatif menurun terjadi berkaitan dengan pelaksanaan vaksinasi pertama pada Presiden Joko Widodo. Berdasarkan hasil clustering tweet positif yang dilakukan pada penelitian ini, dapat disimpulkan bahwa masyarakat mulai menunjukkan partisipasi untuk melakukan vaksin walaupun belum menyeluruh setelah pelaksanaan vaksin yang dilakukan oleh presiden. Penelitian selanjutnya dapat melakukan analisis lebih dalam mengenai perbandingan kinerja beberapa algoritme *Machine Learning* dalam melakukan analisis sentimen pengguna twitter terhadap vaksinasi Covid-19.

DAFTAR PUSTAKA

[BPS] Badan Pusat Statistik. 2020. *Hasil Survei Sosial Demografi Dampak COVID-19 2020.* Jakarta(ID): BPS RI.

Ben-Hur A. 2010. A user's guide to support vector machines. *Methods in Molecular Biology.* Singapore: Springer.

Bholowalia P, Kumar A. 2014. EBK-Means: a clustering technique based on elbow method and KMeans in WSN. *Int J Comput Appl*. 105(9):17-24.

Budiharto W, Meiliana M. 2018. Prediction and analysis of Indonesia presidential election from twitter using sentiment analysis. *Journal of Big Data.* 51(2018).

Dewi RK. 16 Des 2020. Pemerintah gratiskan vaksin covid-19, ini 6 kelompok prioritas vaksinasi [Internet]. [diakses 2021 Jun 1]. https://www.kompas.com/tren/read/2020/12/16/152133765/pemerintah-gratiskan-vaksin-covid-19-ini-6-kelompok-yang-jadi-prioritas.

Diya W, Yixi Z. 2019. Using news to predict investor sentiment: based on SVM model. Di dalam: *2019 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI2019);* Jinan, 2019 Okt 25-27. Procedia Computer Science. hlm 191-199.

Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. 2011. An overview of ensemble methods for binary classifiers in multi-class problems: experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition*. 44(8): 1761-1776.

Gong Z, Cai T, Thill JC, Hale S, Graham M. 2020. Measuring relative opinion from location-based social media: a case study of the 2016 US presidential election. *PLoS ONE*. 15(5): e0233660.

Hidayatullah. 2021. Gerakan tolak vaksin COVID-19, akankah berakhir lewat anjuran MUI dan tokoh agama? [Internet]. [diakses 2021 Jun 1]. https://www.bbc.com/indonesia /indonesia-55644537.

Jia X, Junxiang C, Chen C, Chengda Z, Sijia L, Tingshao Z. 2020. Public discourse and sentimen during the COVID 19 pandemic: using latent dirichlet allocation for topik modelling on Twitter. *PLoS ONE*. 15(9): e0239441.

Joachims T. 1998. Text categorization with support vector machines: Learning with many relevant features. Di dalam: *Proc of the European Conference on Machine Learning (ECML)*. Lecture Notes in Computer Science Vol. 1398. hlm.137–142.

Lema G, Nogueira F, Aridas CK. 2017. A Python toolbox to tackle the curse of imbalanced dataset in machine learning. *Journal of Machine Learning Research*. 18(17): 1-5.

Luqyana WA, Cholissodin I, Perdana RS. 2018. Analisis sentimen *cyberbullying* pada komentar instagram dengan metode klasifikasi *support vector machine*. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Universitas Brawijaya*. 2(11): 4704-4713.

Nugroho A, Witarto A, Handoko D. 2003. Application of support vector machine in bioinformatics. Di dalam: *Proceeding of Indonesian Scientific Meeting in Central*. hlm 19-27.

Nuraini R. 2020. Kasus Covid-19 pertama, masyarakat jangan panik [Internet]. [diakses 2021 Jun 1]. https://indonesia.go.id/narasi/indonesia-dalam-angka/ekonomi/kasus-covid-19-pertama-masyarakat-jangan-panik.

Pang B, Lee L, Vaithyanathan S. 2002. Thumbs up? Sentiment classification using machine learning techniques. Di dalam: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*. hlm 79–86.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, *et al.* 2011. Scikit-learn: machine learning in Python. *JMLR.* 12(85): 2825-2830.

Robertson SE. 2004. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*. 60(5): 503-520.

Shalihah NF. 2020. Kasus dan kematian akibat covid-19 di Indonesia tertinggi di ASEAN. [diakses 2021 Jun 1]. https://www.kompas.com/tren/read/2020/10/16/141000165/kasus-dan-kematian-akibat-covid-19-di-indonesia-tertinggi-di-asean.

Shivaprasad TK, Shetty J. 2017. Sentimen analysis of product reviews: a review. Di dalam: *International Conference on Inventive Communication and Computational Technologies*. New York: IEEE.

Vapnik V, Cortes C. 1995. Support vector networks. *Machine Learning*. 20(1995): 273-297.

Xueting W, Canruo Z, Zidian X, Dongmei L. 2020. Public opinions towards COVID-19 in California and New York on Twitter. *medRxiv*.

Yousefinaghani S, Dara R, Mubareka S, Papadopoulos, Sharif S. 2021. An analysis of COVID-19 vaccine sentiments and opinions on Twitter. *International Journal of Infectious Diseases.* 108(2021): 256-262.