Analisis Sentimen Pelaksanaan Vaksinasi Covid-19 secara Massal pada Media Sosial Twitter

(Sentiment Analysis of Mass Covid-19 Vaccination on Twitter Social Media)

Adinda Febby Nuraini[1*], Rosma Dian Pertiwi[1], Muhammad Zidni Subarkah[1], Kiki Ferawati[1]

[1]*Program Studi Statistika, Universitas Sebelas Maret, Surakarta, Jawa Tengah, Indonesia*
Jalan Ir. Sutami 36 Kentingan, Jebres, Surakarta, Jawa Tengah. Indonesia 57126
E-mail: adindafebby@student.uns.ac.id

ABSTRAK

Virus corona atau Covid-19 menyerang hampir seluruh negara di dunia, begitu juga Indonesia. Pemerintah Indonesia melaksanakan beberapa kebijakan untuk menangani penyebaran virus Covid-19 di masyarakat yaitu dengan vaksinasi massal. Pelaksanaan vaksinasi massal menjadi pembicaraan masyarakat salah satunya di media sosial Twitter. Penelitian ini bertujuan untuk menganalisis sentimen pelaksanaan vaksinasi massal di Indonesia menggunakan perbandingan antara klasifikasi *Naïve Bayes*, *Random Forest*, dan *Support Vector Machine* (SVM). Hasil penelitian menunjukan bahwa penggunaan metode klasifikasi SVM menghasilkan *F1-Score* lebih tinggi daripada metode lain yaitu sebesar 84%. Selain itu, dapat disimpulkan bahwa sebagian besar masyarakat pro terhadap pelaksanaan vaksin massal. Sehingga metode SVM dapat digunakan pemerintah untuk mengklasifikasikan sentimen masyarakat terhadap pengadaan vaksinasi massal selanjutnya dan dijadikan dasar pemerintah untuk mempertahankan program vaksinasi massal ini sebagai upaya mencegah penyebaran Covid-19.

Kata kunci: Vaksinasi Massal, Analisis Sentimen, *Naïve Bayes, Random Forest, Support Vector Machine*

ABSTRACT

The corona virus or Covid-19 attacks almost all countries in the world, as well as Indonesia. The Indonesian government has implemented several policies to deal with the spread of the Covid-19 in the community, namely by mass vaccination. The implementation of mass vaccination has become trending on Twitter. This study aims to analyze the sentiment of mass vaccination in Indonesia using a comparison between Naïve Bayes, Random Forest, and Support Vector Machine (SVM) methods. The results showed that SVM classification method has F1-Score higher than other methods, which was 84%. In addition, it can be concluded that most of the community is pro against the implementation of mass vaccines. So, SVM method can be used by the government to classify public sentiment towards the next mass vaccination and basis for the government to maintain this mass vaccination program as an effort to prevent the spread of Covid-19.

Keywords: *Mass Vaccination, Sentiment Analysis, Naïve Bayes, Random Forest, Support Vector Machine*

PENDAHULUAN

Penyakit Covid-19 yang disebabkan oleh virus SARS-CoV-2 menyebar dengan pesat di berbagai negara sehingga pada 11 Maret 2020 ditetapkan menjadi pandemi oleh *World Health Organization*. Adanya pandemi Covid-19 menyebabkan dampak yang buruk bagi berbagai sektor. Pesatnya penyebaran dan bahaya Covid-19 perlu segera ditangani. Salah satu cara untuk mencegah penyebaran virus ini dengan mengembangkan vaksin (Rachman dan Pramana, 2020). Pada tahun 2020 beberapa negara sudah melakukan vaksinasi, salah satunya Indonesia. Wakil Menteri Kesehatan RI, dr. Dante Saksono Harbuwono mengatakan kebijakan pelaksanaan vaksinasi massal dapat mempercepat pemerintah dalam mencapai target cakupan vaksinasi Covid-19. Vaksinasi massal secara resmi dimulai pada tanggal 13 Januari 2021 (Kementerian Komunikasi dan Informasi Republik Indonesia, 2021).

Pembahasan mengenai vaksinasi massal sempat menjadi *trending* di media sosial Twitter terutama pada awal pelaksanaannya. Hal ini dikarenakan munculnya masalah mengenai proses pelaksanaan vaksinasi massal Covid-19. Masalah yang ada memunculkan berbagai tanggapan dari masyarakat mengenai pelaksanaan vaksinasi baik tanggapan positif maupun negatif. Pendapat-pendapat yang ada di Twitter dapat dimanfaatkan untuk melakukan analisis sentimen terhadap vaksinasi Covid-19 dengan cara mengklasifikasikan pendapat dan opini ke dalam dua kelas yaitu negatif dan positif (Fitriana dkk., 2021).

Analisis sentimen dapat dilakukan menggunakan beberapa metode seperti *Naïve Bayes, Random Forest,* dan *Support Vector Machine* (SVM). Penelitian terkait mengenai analisis sentimen menggunakan metode *Naïve Bayes* adalah penelitian yang berjudul *"Analisis Sentimen Aplikasi E-Government Pada Google Play Menggunakan Algoritma Naïve Bayes"* oleh Tanggraeni dkk. Pada penelitian ini aplikasi *e-government* pada

Google Play menggunakan algoritma *Naïve Bayes* diperoleh akurasi dengan pembobotan TF-IDF sebesar 89% (Tanggraeni dkk., 2022). Metode *Random Forest* lebih banyak diterapkan untuk klasifikasi dalam beberapa tahun terakhir sejak kinerja dari jenis algoritma ini melampaui SVM, *Naïve Bayes* dan algoritma pembelajaran mesin lainnya. Penelitian terkait mengenai analisis sentimen menggunakan metode *Random Forest* adalah penelitian yang berjudul *"An Evaluation of Preprocessing Steps and Tree-based Ensemble Machine Learning for Analysing Sentiment on Indonesia Youtube Comments"* oleh Aribowo dkk. Pada penelitian Aribowo dkk (2020) memperoleh akurasi sebesar 86,76%. Sedangkan penelitian Ilmawan dan Winarko yang berjudul "*Aplikasi Mobile untuk Analisis Sentimen pada Google Play"* menggunakan metode SVM menghasilkan akurasi yang lebih besar dibandingkan metode *Naïve Bayes* dengan akurasi sebesar 89,49%. Selain itu, metode SVM cepat dalam mengklasifikasi dan bisa mentoleransi atribut yang tidak relevan (Ilmawan dan Winarko, 2015). Kemudian penelitian terkait analisis sentimen mengenai Vaksinasi Covid-19 adalah penelitian yang berjudul Pada jurnal *"Analisa Sentimen Vaksinasi Covid-19 dengan Metode Support Vector Machine dan Naive Bayes Berbasis Teknik Smote"* dari Fahlapi dan teman-teman. Di dalam penelitian tersebut dilakukan analisis sentimen mengenai vaksinasi Covid-19 dengan satu kata kunci yaitu "vaksin". Metode yang digunakan oleh Fahlapi dkk adalah SVM dan *Naïve Bayes* dengan teknik SMOTE dan didapatkan metode terbaik yaitu SVM dengan akurasi 70.51%.

Berdasarkan latar belakang di atas, dilakukan penelitian mengenai analisis sentimen pengguna Twitter terhadap pelaksanaan vaksinasi massal di Indonesia dengan membandingkan metode *Naïve Bayes*, *Random Forest,* dan SVM. Jika dibandingkan dengan penelitian dari Fahlapi dan teman-teman penelitian ini menambahkan beberapa kata kunci agar bisa lebih akurat serta menggunakan tiga metode klasifikasi. Penelitian ini penting dilakukan agar dapat memberikan gambaran mengenai respon pengguna Twitter terhadap pelaksanaan vaksinasi Covid-19 secara massal di Indonesia sehingga dapat membantu pemerintah untuk menentukan kebijakan mengenai vaksinasi massal ke depannya. Selain itu penelitian ini dilakukan dengan harapan mendapatkan suatu metode yang tepat untuk dapat mengklasifikasikan sentimen pada twitter mengenai vaksinasi Covid-19 secara massal

METODE

Crawling Data

Twitter merupakan salah satu media sosial paling populer yang berisi percakapan berupa "*tweet*" di antara sekelompok orang. Di dalamnya, Twitter mengizinkan pengunduhan data dari masing-masing SNS dengan izin terbatas menggunakan perangkat lunak *Application Programming Interface* (API) (Sohail *et al*., 2021). Pada API ini terdapat beberapa kasus penggunaan populer diantaranya untuk menganalisis percakapan sebelumnya, memperoleh *tweet* berdasarkan timeline seseorang, dan analisis sentimen di penambangan teks.

Penambangan teks merupakan penggalian data berupa kata-kata yang diperoleh dari data tidak terstruktur pada kalimat-kalimat dalam suatu dokumen (Maulana and Pratiwi, 2019). Penambangan teks dapat dilakukan dengan cara *crawling* data. *Crawling* adalah teknik yang digunakan untuk mengumpulkan data yang terdapat dalam suatu web. Di mana akan diperoleh informasi sesuai dengan kata kunci yang dicari dengan bantuan Twitter API (Dikiyanti, Rukmi and Irawan, 2021).

Analisis Sentimen

Analisis sentimen didefinisikan sebagai machine learning untuk mengekstrak informasi subjektif dan opini terkait emosi, sikap, suasana hati dan yang lainnya guna menghitung polaritas yang diungkapkan oleh teks tersebut (Sarkar, 2019). Setiap teks ini akan melalui proses pemberian label untuk menentukan tweet tersebut termasuk ke dalam sentimen positif atau negatif. Sentimen positif berisi pujian, saran, suasana hati seperti senang, dan sikap mendukung lainnya. Sedangkan sentimen negatif berisi keluhan, sindiran, amarah, kecewa, dan sikap-sikap kontra (Samsir dkk.., 2021).

Preprocessing Data

Algoritma *supervised* atau *unsupervised* dalam *machine learning* perlu melakukan *preprocessing* sebelum menganalisis data (Maulana and Pratiwi, 2019). Dalam *preprocessing text* terdiri *dari case folding, stopwords removal, tokenization, punctuation removal,* dan *stemming* (Mining, 2020). *Case holding* merupakan suatu proses mengubah huruf menjadi bentuk yang sama, seperti pengubahan ke huruf kecil. *Punctuation removal* adalah suatu proses dalam menghilangkan tanda baca. Selanjutnya, *stopwords removal* digunakan untuk menghapus kata-kata

yang tidak penting. *Tokenization* digunakan untuk membagi kalimat menjadi beberapa bagian yang disebut token. Sedangkan *stemming* berguna untuk mengubah menjadi bentuk kata dasar.

Model *Bag-of-words*

Dalam model *bag-of-words*, teks dalam bentuk kalimat ataupun dokumen direpresentasikan sebagai kantung (bag) mengenai informasi-informasi penting tanpa mengurutkan setiap katanya dan tetap mempertahankan keragamannya (Trisari dan Retno, 2020). Proses mengubah teks menjadi vektor ini digunakan untuk menghitung frekuensi muncul suatu token tertentu.

Pembobotan TF-IDF

Pembobotan TF-IDF (Term Frequency-Inverse Document Frequency) merupakan proses dalam transformasi data tekstual ke dalam numerik yang berguna untuk memberikan pembobotan pada setiap fitur. TF-IDF adalah ukuran statistik untuk mengevaluasi seberapa penting sebuah kata. TF merupakan frekuensi kemunculan kata di dalam dokumen. DF merupakan frekuensi dokumen yang mengandung kata tersebut. IDF merupakan *inverse* dari DF. Hasil dari pembobotan kata TF-IDF merupakan hasil dari perkalian TF yang dikalikan dengan IDF. Kata yang sering muncul maka akan memiliki bobot yang lebih besar dalam suatu dokumen dan akan semakin kecil jika muncul di dalam banyak dokumen (Septian, Fachrudin and Nugroho, 2019). Perhitungan TF-IDF dapat melalui persamaan berikut:

$$w_{i.j} = tf_{i.j} \times idf_j = tf_{ij} \times log(\frac{N}{df_j}) \quad \text{..... (1)}$$

dengan:
$w_{i.j} =$ bobot *term* (t_j) terhadap dokumen (d_i)
$tf_{i.j} =$ jumlah kemunculan *term* (t_j) terhadap dokumen (d_i)
$idf_j =$ *inverse document term* (t_j)
$N =$ jumlah dokumen
$df_j =$ jumlah dokumen yang mengandung *term* (t_j)

Naïve Bayes

Algoritma *Naïve Bayes* merupakan algoritma *supervised learning* yang menerapkan teorema *Bayes* dan terdapat asumsi 'naif' yang menyatakan bahwa setiap fitur independen satu sama lain. Setiap fitur dianggap berkontribusi pada probabilitas setiap pengujian. Sehingga dengan menggunakan teorema *Bayes* kita dapat menentukan probabilitas suatu variabel respon (Sarkar, 2019). Metode naïve bayes merupakan metode yang memiliki perhitungan matematik dasar yang sangat kuat dan stabil, namun memiliki kekurangan yaitu parameter model perlu diperkirakan dan kurang peka dengan data hilang. Model ini mempunyai tingkat kesalahan yang sangat minimum dan metode ini juga populer untuk kasus klasifikasi pada teks (Pratiwi, Yusuf and Nugroho, 2016).

Random Forest

Random Forest merupakan algoritma *ensemble learning* yang pengklasifikasiannya berdasarkan hasil yang diperoleh dari sekumpulan *decision tree* (Nayak, 2016). Banyaknya pohon dapat mengurangi varians dalam model keseluruhan dan mengendalikan *overfitting*. Serta semakin banyak pohon di *Random Forest* semakin tinggi hasil akurasi yang diberikan (Fauzi, 2018). Metode ini adalah pengembangan dari metode *Classification and Regression Tree* (CART) yang menerapkan *random feature selection* dan *bootstrap aggregating*. Kelebihan metode ini yaitu dapat menghasilkan akurasi yang lebih tinggi dalam jumlah data yang besar (Ramadhan, Susetyo and Indahwati, 2019).

Support Vector Machine

Support Vector Machine (SVM) adalah *algoritma supervised learning* yang digunakan untuk klasifikasi, regresi, dan deteksi anomali atau *outlier*. Dalam menerapkan model SVM ini membutuhkan suatu parameter. Di mana keakuratan model sangat bergantung pada parameter tersebut, dan karenanya nilai optimal akan didapat dari teknik *grid-search* yang sesuai (Nayak, 2016). SVM adalah suatu teknik untuk menangani kasus seperti klasifikasi. SVM memiliki dasar *linier classifier* dan telah dikembangkan untuk masalah *non-linier*. SVM menggunakan konsep kernel dengan dimensi tinggi. Pada ruang berdimensi tinggi, akan ditemukan *hyperplane* untuk memaksimalkan jarak antara *class data* (Octaviani, Yuciana Wilandari and Ispriyanti, 2014)

Evaluasi Model

Pada tahap evaluasi model, untuk mengetahui akurasi algoritma dilakukan evaluasi akurasi, presisi, dan *recall* untuk menghasilkan kesimpulan performa dari algoritma yang dipakai (Fitri dkk., 2019) untuk evaluasi ini digunakan *confusion matrix* seperti pada Tabel 1.

Tabel 1. Evaluasi Model Klasifikasi

Klasifikasi	Positif (prediksi)	Negatif (prediksi)
Positif (aktual)	*True* Positif (TP)	*False* Negatif (FN)
Negatif (aktual)	*False* Positif (FP)	*True* Negatif (TN)

Tingkat akurasi dapat dihitung melalui persamaan berikut:

$$Akurasi = \frac{TP+TN}{TP+FN+FP+TN} \quad (2)$$

Presisi adalah jumlah prediksi yang dibuat yang benar-benar relevan dari semua prediksi berdasarkan kelas positif. Semakin tinggi presisi maka akan mengembalikan hasil yang lebih relevan. Presisi dapat dihitung melalui persamaan berikut:

$$Presisi = \frac{TP}{TP+FP} \quad (3)$$

Berbeda dengan presisi, *recall* merupakan jumlah dari kelas positif yang diprediksi dengan benar yang dihitung melalui persamaan berikut:

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

Selain itu, dalam mengevaluasi model juga bisa menggunakan F1-*Score* atau F-*measure* yang bisa dihitung menggunakan persamaan berikut :

$$F1 - Score = 2 * \frac{Recall*Presisi}{Recall+Presisi} \quad (5)$$

Data dan Sumber Data

Pengumpulan data menggunakan teknik *crawling* melalui API Twitter dengan memanfaatkan paket *tweepy*. Data yang diambil berdasarkan kata kunci yang sudah ditentukan dalam periode 7 Juli 2021 hingga 14 Juli 2021 dan selanjutnya disimpan dalam bentuk CSV.

Langkah Analisis

Proses analisis menggunakan bahasa pemrograman *Python* yang terdiri dari 4 tahap secara garis besar yaitu pengumpulan data, *preprocessing* data, eksplorasi data, dan proses klasifikasi serta evaluasi. Klasifikasi menggunakan metode *Naïve Bayes, Random Forest*, dan *Support Vector Machine.* Berikut ini tahapan-tahapan yang dilakukan :

A. Pengumpulan Data

1. Pengumpulan data menggunakan teknik *crawling* melalui API Twitter. Data yang diambil berdasarkan kata kunci yang berhubungan dengan vaksinasi Covid-19 secara massal yaitu vaksin massal, antre vaksin, kerumunan vaksin, vaksinasi massal, vaksinasi serentak, vaksin serentak, vaksin covid massal, dan imunisasi covid massal dalam periode 7 Juli 2021 hingga 14 Juli 2021 yang selanjutnya disimpan dalam bentuk CSV.
2. Pelabelan data *tweet* dilakukan secara manual oleh tiga peneliti dengan penentuan label menggunakan metode modus hasil pelabelan peneliti. Metode ini bertujuan untuk mengurangi bias hasil penelitian dari label sentimen. Pelaksanaan vaksinasi massal menimbulkan pro dan kontra di tengah masyarakat, sehingga pelabelan hanya dikelompokkan ke dalam dua kelas yaitu sentimen positif dan negatif. Pelabelan sentimen positif berdasarkan *tweet* yang berisi tentang dukungan vaksinasi massal. Sedangkan pelabelan sentimen negatif berisi tentang kritikan atau penolakan terkait adanya vaksinasi massal.

***B. Preprocessing* Data**

1. Menghapus data *tweet* yang tidak terdapat topik mengenai vaksinasi massal di Indonesia dan data *tweet* yang tidak mengarah ke sentimen positif atau negatif.
2. Membersihkan data dengan menghapus tautan HTML di *tweet*, emoji, *mention*, *username*, *hashtag*, "RT" atau *retweet*, angka, dan tanda baca, menghapus *strip* ruang putih seperti spasi, tab, baris baru yang berlebihan.
3. Menghapus data yang terduplikat.
4. Melakukan *case folding* yaitu mengubah semua huruf menjadi tidak kapital dan melakukan *filtering* dengan menghapus kata-kata dalam *tweet* yang terdapat dalam *stopwords* berdasarkan paket Sastrawi.
5. Melakukan tokenisasi yaitu memotong kalimat menjadi kata-kata.

6. Melakukan *feature extraction* dengan *Bag of Words* (BoW) dan melakukan pembobotan menggunakan TF-IDF.
7. Membagi data menjadi data latih dan data uji dengan perbandingan 80:20 di mana data latih sebanyak 80% dan data uji adalah 20%.

C. Eksplorasi Data

Eksplorasi data bertujuan untuk mengetahui deskripsi dan visualisasi data menggunakan *word cloud* berdasarkan *tweet* positif dan negatif.

D. Pengklasifikasian dan Evaluasi

Tahap klasifikasi data dilakukan dengan tiga metode yaitu *Naïve Bayes, Random Forest*, dan *Support Vector Machine.* Untuk tahap validasi menggunakan 10 fold *Cross Validation*. Mengevaluasi metode klasifikasi dengan *F1-Score* setiap kelas dan akurasi setiap metode kemudian membandingkan antara tiga metode tersebut.

HASIL DAN PEMBAHASAN

Pada hasil dan pembahasan ini peneliti menyajikan proses analisis yang telah dilakukan. Penelitian ini mengulas tentang sentimen vaksin massal di Indonesia, dimana dari beberapa penelitian sebelumnya tidak ditemukan bahwa topik vaksin massal ini telah dilakukan penelitian sebelumnya. Penelitian ini merupakan penelitian baru yang dapat menyajikan *insight* yang berguna dan bermanfaat. Peneliti melakukan analisis sentimen menggunakan beberapa metode yaitu dengan membandingkan metode *Naïve Bayes, Random Forest,* dan SVM.

Pengumpulan Data

Setelah melakukan *crawling* dari Twitter, berdasarkan kata kunci sebanyak 6.000 *tweet* dan didapatkan 1.461 *tweet* setelah dibersihkan *tweet* duplikat. Langkah selanjutnya adalah memberi label dengan mengambil modus label dari pelabelan peneliti, di mana terdapat dua kelas yaitu sentimen positif dan negatif. Data *tweet* yang tidak memiliki sentimen dan yang tidak terkait dengan vaksin massal tidak digunakan. Tabel 2 menunjukkan beberapa sampel data sentimen positif dan negatif.

Tabel 2. Beberapa Sampel Data *Tweet*

No.	*Tweet*	Label
1.	Seneng banget sekarang, akhirnya ngga cuman orang tua diatas 60 tahun dan masyarakat diatas 18 tahun yang mendapat vaksinasi. Program vaksinasi massal menyerbu sekolah untuk umur 12-17 tahun juga. Hayoo siapa yang belum vaksin, yuk segera cari informasi #GerakanBerbagiUntukWarga https://t.co/vdmTizS1Dx	Positif
2.	@uusbiasaaja Nakes : "kamu kena covid abis kerumunan darimana?" Orang : "kerumunan antri vaksin dok" HaduhaduhðŸ˜,	Negatif

Preprocessing Data

Tujuan dari *preprocessing* pada data adalah untuk menghilangkan noise, memperjelas fitur, meningkatkan akurasi, dan mengkonversi data asli sehingga dapat diolah sesuai dengan kebutuhan penelitian. Berikut adalah beberapa langkah dalam preprocessing data:

1. Menghapus data *tweet* yang tidak terkait dengan vaksinasi massal di Indonesia dan *tweet* yang tidak mengandung sentimen positif atau negatif.
2. Pembersihan *tweet* :
   a. Hapus tautan HTML di setiap *tweet* dan ikon emoji

Dalam data *tweet* terkadang terdapat tautan HTML dan ikon emoji yang dapat mengganggu akurasi model. Tabel 3 merupakan proses penghapusan tautan HTML dan ikon emoji.

Tabel 3. *Tweet* Setelah di Hapus Tautan HTML dan Ikon Emoji

No.	Sebelum	Sesudah
1.	Seneng banget sekarang, akhirnya ngga cuman orang tua diatas 60 tahun dan masyarakat diatas 18 tahun yang mendapat vaksinasi. Program vaksinasi massal menyerbu sekolah untuk umur 12-17 tahun juga. Hayoo siapa yang belum vaksin, yuk segera cari informasi	Seneng banget sekarang, akhirnya ngga cuman orang tua diatas 60 tahun dan masyarakat diatas 18 tahun yang mendapat vaksinasi. Program vaksinasi massal menyerbu sekolah untuk umur 12-17 tahun juga. Hayoo siapa yang belum vaksin, yuk segera cari informasi

Tabel 3 (lanjutan)

No.	Sebelum	Sesudah
	#GerakanBerbagiUntukWarga https://t.co/vdmTizS1Dx	#GerakanBerbagiUntukWarga
2.	@uusbiasaaja Nakes : "kamu kena covid abis kerumunan darimana?" Orang : "kerumunan antri vaksin dok" HaduhaduhðŸ˜,	@uusbiasaaja Nakes : "kamu kena covid abis kerumunan darimana?" Orang : "kerumunan antri vaksin dok" Haduhaduh

b. Hapus sebutan, nama pengguna, tagar, angka, "RT" atau *retweet,* tanda baca, dan *Strip* Ruang Putih

Tabel 4. *Tweet* setelah di Hapus Sebutan, Nama Pengguna, Tagar, Angka, "RT" atau *retweet,* Tanda Baca, dan *Strip* Ruang Putih

No.	Sebelum	Sesudah
1.	Seneng banget sekarang, akhirnya ngga cuman orang tua diatas 60 tahun dan masyarakat diatas 18 tahun yang mendapat vaksinasi. Program vaksinasi massal menyerbu sekolah untuk umur 12-17 tahun juga. Hayoo siapa yang belum vaksin, yuk segera cari informasi #GerakanBerbagiUntukWarga	Seneng banget sekarang akhirnya ngga cuman orang tua diatas tahun dan masyarakat diatas tahun yang mendapat vaksinasi Program vaksinasi massal menyerbu sekolah untuk umur tahun juga Hayoo siapa yang belum vaksin yuk segera cari informasi
2.	@uusbiasaaja Nakes : "kamu kena covid abis kerumunan darimana?" Orang : "kerumunan antri vaksin dok" Haduhaduh,	Nakes kamu kena covid abis kerumunan darimana Orang kerumunan antri vaksin dok Haduhaduh

3. Menghapus data duplikat

Setelah melakukan *drop* duplikat diperoleh sebanyak 1.320 *tweet* dari 1.421 *tweet* berisi sentimen positif atau negatif terkait vaksinasi massal di Indonesia.

4. *Case folding* dan filtering dengan *stopword*

Daftar *stopwords* adalah kumpulan kata-kata umum yang biasanya muncul dalam jumlah besar dan dianggap tidak ada artinya. Fungsi dari penghapusan kata berdasarkan daftar *stopwords* adalah untuk mengurangi jumlah kata dan menghilangkan *noise* seperti pada Tabel 5.

Tabel 5. *Tweet* Setelah di *Case folding* dan Hapus *Stopword*

No.	Sebelum	Sesudah
1.	Seneng banget sekarang akhirnya ngga cuman orang tua diatas tahun dan masyarakat diatas tahun yang mendapat vaksinasi Program vaksinasi massal menyerbu sekolah untuk umur tahun juga Hayoo siapa yang belum vaksin yuk segera cari informasi	seneng sekarang akhirnya orang tua diatas tahun masyarakat diatas tahun mendapat vaksinasi program vaksinasi massal menyerbu sekolah umur tahun siapa belum vaksin yuk segera cari informasi
2.	Nakes kamu kena covid abis kerumunan darimana Orang kerumunan antri vaksin dok Haduhaduh	nakes kamu kena covid abis kerumunan orang kerumunan antri vaksin dok

5. Tokenisasi

Tabel 6. *Tweet* Setelah di Tokenisasi

No.	Sebelum	Sesudah
1.	seneng sekarang akhirnya orang tua diatas tahun masyarakat diatas tahun mendapat vaksinasi program vaksinasi massal menyerbu sekolah umur tahun siapa belum vaksin yuk segera cari informasi	“seneng”, “sekarang”, “akhirnya”, “orang”, “tua”, “diatas”, “tahun”, “masyarakat”, “diatas”, “tahun”, “mendapat”, “vaksinasi”, “program”, “vaksinasi”, “massal”, “menyerbu”, “sekolah”, “umur”, “tahun”, “siapa”, “belum”, “vaksin”, “yuk”, “segera”, “cari”, “informasi”
2.	nakes kamu kena covid abis kerumunan orang kerumunan antri vaksin dok	“nakes”, “kamu”, “kena”, “covid”, “abis”, “kerumunan”, “orang”, “kerumunan”, “antri”, “vaksin”, “dok”

6. Melakukan *bag-of-words* (BoW) dan Pembobotan dengan TF-IDF

Dalam penelitian ini menggunakan model *bag-of-words* (BoW) di mana berdasarkan sampel data *tweet*, maka sebuah list dibentuk untuk tiap dokumen pada Tabel 7.

Tabel 7. *List* dalam Bentuk Dokumen

Dokumen	Tweet	Tiap Dokumen
1	seneng sekarang orang orang vaksin	"seneng", "sekarang", "orang", "orang", "vaksin"
2	nakes kena covid abis antri vaksin	"nakes", "kena", "covid", "abis", "antri", "vaksin"

Selanjutnya untuk setiap kata dihitung frekuensinya dan dipetakan kembali ke dalam dokumen pada Tabel 8.

Tabel 8. Frekuensi Kata Dalam Dokumen

	Dokumen	seneng	Sekarang	orang	vaksin	nakes	kena	covid	abis	antri
1	seneng sekarang orang orang vaksin	1	1	2	1	0	0	0	0	0
2	nakes kena covid abis antri vaksin	0	0	0	1	1	1	1	1	1

Berdasarkan corpus atau kumpulan kata yang dibentuk dari tiap dokumen, dihitung pembobotan menggunakan TF-IDF dengan hasil pada Tabel 9.

Tabel 9. Perhitungan Pembobotan TF-IDF

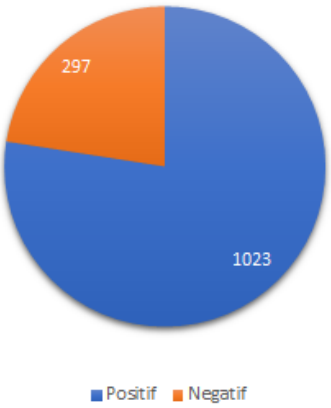
	Dokumen	seneng	Sekarang	orang	vaksin	nakes	kena	covid	abis	antri
1	seneng sekarang orang orang vaksin	1.405	1.405	2.811	1	0	0	0	0	0
2	nakes kena covid abis antri vaksin	0	0	0	1	1.405	1.405	1.405	1.405	1.405

Evaluasi untuk mengetahui seberapa penting sebuah kata atau istilah dalam kumpulan dokumen atau *corpus* juga merupakan salah satu tahap dalam penambangan teks dan pengambilan informasi. Hal ini ditekankan kata-kata dalam dokumen yang sering muncul di *corpus*.

Eksplorasi

1. Statistik Deskriptif

Selanjutnya melakukan analisis statistik deskriptif dengan tujuan untuk melihat pada eksplorasi data secara keseluruhan. Gambar 1 menunjukkan diagram lingkaran mengenai proporsi sentimen positif dan negatif.



Gambar 1. Diagram Lingkaran Label Sentimen

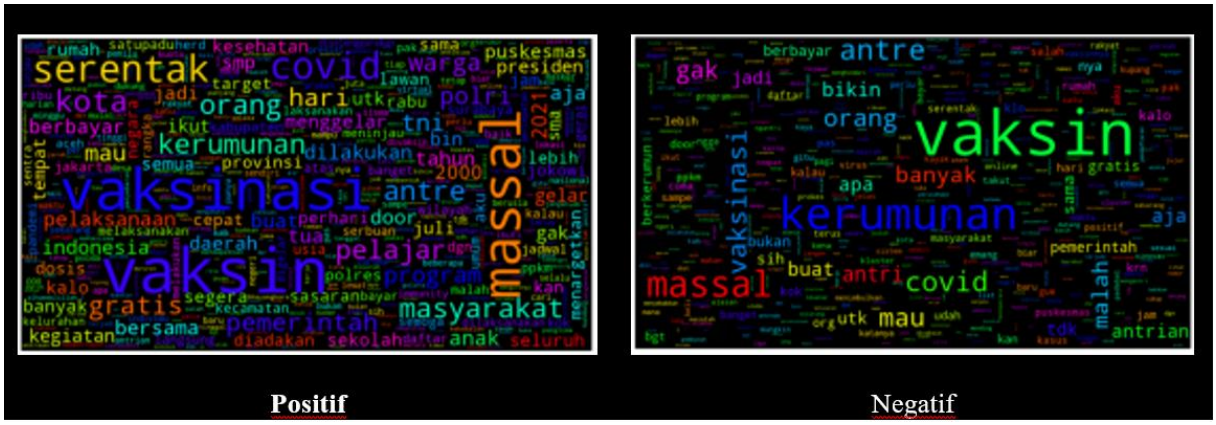
Berdasarkan Gambar 1 di atas terlihat bahwa *tweet* yang mengandung sentimen positif sebanyak 1.023 *tweet* sedangkan sentimen negatif 297 *tweet*. Terlihat bahwa sentimen positif lebih banyak dari sentimen negatif. Hal ini menunjukan bahwa jumlah label dalam *tweet* mengalami *unbalance data* sehingga jika dilakukan analisis maka analisis akan lebih memihak pada data yang banyak. Ada beberapa solusi yang perlu diterapkan dalam mengatasi kasus ini, salah satu solusi yang bisa diterapkan adalah dengan menggunakan metrik evaluasi eror *F1-Score.*

Evaluasi eror ini dapat menghasilkan evaluasi yang dapat lebih dipertanggungjawabkan karena *F1-Score* merupakan interpretasi dari evaluasi eror pada setiap sentimen.

2. Visualisasi data dengan *word cloud*

Pada Gambar 2 *word cloud* sentimen positif didapatkan tiga kata yang frekuensinya muncul paling banyak selain *keyword* yaitu, "kota", "orang", dan "masyarakat" dengan masing-masing secara berurutan muncul sebanyak 148, 124, dan 124 frekuensi. Hal ini menunjukkan bahwa "kota" adalah salah satu kata kunci pemicu sentimen positif.

Gambar 2 *word cloud* sentimen negatif didapatkan tiga kata yang frekuensinya muncul paling banyak selain *keyword* yaitu, "kerumunan", "mau", dan "orang" dengan masing-masing secara berurutan muncul sebanyak 168, 53, dan 51 frekuensi. Hal ini menunjukkan bahwa "kerumunan" adalah salah satu kata kunci pemicu sentimen negatif.



Gambar 2. *Word Cloud* Sentimen Positif dan Negatif

Klasifikasi dan Evaluasi Model

Setelah melakukan transformasi, maka dilakukan penggabungan model BoW, TF-IDF, dan metode klasifikasi mengunakan *pipeline*. Kemudian dipilih parameter yang optimum dengan BoW *ngram range* (1,1) dan (1,2) serta *tifidf_use_idf* menggunakan *True* dan *False*. Pada metode *Naïve Bayes Classifier* memiliki tambahan parameter yaitu *classifier_alpha* yaitu (0.01, 0.001), sedangkan metode *random forest* memiliki tambahan *classifier_criterion gini* dan *entropy*. Selanjutnya, untuk metode *Support Vector Machine* menggunakan kernel *linear*. Untuk mengoptimumkan model juga dilakukan *GridSearchCV* dengan kombinasi 10-*fold cross validation* dan *verbose* 1 seperti pada Tabel 10.

Tabel 10. Perbandingan Evaluasi Hasil Klasifikasi

Metode	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Naïve Bayes	0.82	0.63	0.57	81%
Random Forest	0.84	0.76	0.46	82%
Support Vector Machine	0.84	0.70	0.60	84%

Pada Tabel 10 menunjukkan bahwa SVM dan *Random Forest* memiliki akurasi tertinggi. Namun *Precision* ditempati oleh *Random Forest* sedangkan *Recall* ditempati oleh SVM. Selain itu nilai tertinggi pada *F1-Score* yaitu pada SVM. Hal ini menunjukkan bahwa memang perlu adanya peninjauan terhadap nilai *F1-Score*. Berdasarkan tahap eksplorasi data diketahui bahwa data *unbalance* dimana jumlah sentimen positif berbeda jauh dari jumlah sentimen negatif. Oleh karena itu pada penelitian ini untuk membandingkan ketiga metode klasifikasi digunakan *F1-Score*. Penelitian ini tidak menggunakan presisi dan *recall* karena evaluasi *F1-Score* merupakan operasi gabungan dari presisi dan *recall*, sehingga kedua evaluasi tersebut bisa diwakilkan oleh *F1-Score*. Evaluasi *F1-Score* dapat digunakan untuk melihat prediksi kelas positif dan negatif. Sedangkan untuk akurasi tidak digunakan sebagai bahan evaluasi karena data yang digunakan *unbalance.*

KESIMPULAN

Berdasarkan analisis sentimen vaksinasi massal di Indonesia diperoleh bahwa klasifikasi menggunakan metode *Support Vector Machine* (SVM) lebih baik dengan *F1-Score* sebesar 84%, sedangkan metode *Naïve Bayes*

sebesar 81% dan Random Forest sebesar 82%. Oleh karena itu, SVM dapat digunakan pemerintah untuk mengklasifikasikan sentimen masyarakat terhadap pengadaan vaksinasi massal selanjutnya. Selain itu, diperoleh sentimen terbanyak adalah sentimen positif yang mana menunjukkan pelaksanaan vaksinasi massal mendapatkan respon positif atau dukungan dari masyarakat. Untuk penelitian selanjutnya perlu dicoba mengembangkan dengan metode lain dengan data yang lebih banyak dan *real time* untuk mendapatkan hasil akurasi yang lebih baik. Selain itu juga dengan cara mengembangkan daftar kata dalam stopword.

DAFTAR PUSTAKA

Aribowo, A. S., Basiron, H., Herman, N. S., dan Khomsah, S. (2020) 'An evaluation of preprocessing steps and tree-based ensemble machine learning for analyzing sentiment on Indonesian youtube comments', *Int. J. Adv. Trends Comput. Sci. Eng.*, 9(5), pp. 7078–7086. doi: 10.30534/ijatcse/2020/29952020.

Dikiyanti, T. D., Rukmi, A. M. and Irawan, M. I. (2021) 'Sentiment analysis and topic modeling of BPJS Kesehatan based on twitter crawling data using Indonesian Sentiment Lexicon and Latent Dirichlet Allocation algorithm', *Journal of Physics: Conference Series*, 1821(1). doi: 10.1088/1742-6596/1821/1/012054.

Fahlapi, R., Hermanto, Asra, T., Kuntoro, A. Y., Ocanitra, R., Effendi, L., Syukmana., F. (2022) 'Analisa Sentimen Vaksinasi Covid-19 dengan Metode Support Vector Machine dan Naive Bayes Berbasis Teknik SMOTE', *Jurnal Informatika Kaputama (JIK)*, 6(1), pp. 57-63.

Fauzi, M. A. (2018) 'Random forest approach for sentiment analysis in Indonesian language', *Indonesian Journal of Electrical Engineering and Computer Science*, 12(1), pp. 46–50. doi: 10.11591/ijeecs.v12.i1.pp46-50.

Fitri, V. A., Andreswari, R. and Hasibuan, M. A. (2019) 'Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm', *Procedia Computer Science*, 161, pp. 765–772. doi: 10.1016/j.procs.2019.11.181.

Fitriana, F., Utami, E. dan Fatta, H. Al (2021) 'Analisis Sentimen Opini Terhadap Vaksin Covid-19 pada Media Sosial Twitter Menggunakan Support Vector Machine dan Naive Bayes', 5(1), pp. 19–25.

Ilmawan, Lutfi, Budi dan Edi, Winarko. (2015) 'Aplikasi Mobile untuk Analisis Sentimen pada Google Play', *IJCSS*, 9(1), pp 53-64.

Kementerian Kesehatan Republik Indonesia. (2021). *Sentra Vaksinasi Massal Percepat Capai Target Vaksinasi.* Available Online at https://www.kemkes.go.id/article/view/21032000003/sentra-vaksinasi-massal-percepat-capai-target-vaksinasi.html . Accessed 8 July 2022.

Kementerian Komunikasi dan Informasi Republik Indonesia. (2021). *Diawali Presiden, Vaksinasi Covid-19 Gratis Resmi Dimulai di seluruh Indonesia.* Available Online at https://www.kominfo.go.id/content/detail/32066/diawali-presiden-vaksinasi-covid-19-gratis-resmi-dimulai-di-seluruh-indonesia/0/berita. Accessed 28 July 2021.

Maulana, A. and Pratiwi, H. (2019) 'Sentiment analysis of public towards infrastructure development in Indonesia on Twitter media using boosting support vector machine method', *AIP Conference Proceedings*, 2202(2019). doi: 10.1063/1.5141695.

Mining, A. O. (2020) 'Ringkasan Jumlah Aspek Ulasan Hotel untuk Pembentukan Dataset Sentimen Analisis Berbasis Aspek', 3(2), pp. 62–66.

Nayak, A. (2016) 'Comparative study of Naïve Bayes , Support Vector Machine and Random Forest Classifiers in Sentiment Analysis of Twitter feeds', *International Journal of Advanced Studies in Computer Science and Engineering*, 5(1), pp. 14–17.

Octaviani, P. A., Yuciana Wilandari and Ispriyanti, D. (2014) 'Penerapan Metode Klasifikasi Support Vector Machine (SVM) pada Data Akreditasi Sekolah Dasar (SD) di Kabupaten Magelang', *Jurnal Gaussian*, 3(8), pp. 811–820. Available at: http://download.portalgaruda.org/article.php?article=286497&val=4706&title=PENERAPAN METODE KLASIFIKASI SUPPORT VECTOR MACHINE (SVM) PADA DATA AKREDITASI SEKOLAH DASAR (SD) DI KABUPATEN MAGELANG.

Pandemic, C.-, Laurensz, B. dan Sediyono, E. (2021) 'Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19 (Analysis of Public Sentiment on Vaccination in Efforts to Overcome the', 10(2), pp. 118–123.

Pratiwi, R. W., Yusuf, D. and Nugroho, S. (2016) 'Prediksi Rating Film Menggunakan Metode Naïve Bayes', *Jurnal Teknik Elektro*, 8(2), pp. 60–63. Available at: https://www.kaggle.com/.

Rachman, F. F. dan Pramana, S. (2020) 'Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter', *Health Information Management Journal*, 8(2), pp. 100–109.

Available at: https://inohim.esaunggul.ac.id/index.php/INO/article/view/223/175.

Ramadhan, A., Susetyo, B. and Indahwati, I. (2019) 'Penerapan Metode Klasifikasi Random Forest Dalam Mengidentifikasi Faktor Penting Penilaian Mutu Pendidikan', *Jurnal Pendidikan dan Kebudayaan*, 4(2), p. 169. doi: 10.24832/jpnk.v4i2.1327.

Rathan, M. Vishwanath R. H., Venugopal K. R., and Patnaikc L. M. (2018) 'Consumer insight mining: Aspect based Twitter opinion mining of mobile phone reviews', *Applied Soft Computing Journal*, 68, pp. 765–773. doi: 10.1016/j.asoc.2017.07.056.

Samsir *et al.* (2021) 'Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes', *Jurnal Media Informatika Budidarma*, 5, pp. 157–163. doi: 10.30865/mib.v5i1.2604.

Sarkar, D. (2019) *Text Analytics with Python*, *Text Analytics with Python*. doi: 10.1007/978-1-4842-4354-1.

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', *Journal of Intelligent System and Computation*, 1(1), pp. 43–49. doi: 10.52985/insyst.v1i1.36.

Sohail, S. S. *et al.* (2021) 'Crawling Twitter data through API: A technical/legal perspective'. Available at: http://arxiv.org/abs/2105.10724.

Tanggraeni, A. I. dan Sitokdana, M. N. N. (2022) 'Analisis Sentimen Aplikasi E-Government pada Google Play Menggunakan Algoritma Naïve Bayes', *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, 9(2), pp. 785–795. doi: 10.35957/jatisi.v9i2.1835.

Trisari W. H. P. dan Retno H. (2020) 'Penggalian Teks Dengan Model Bag of Words Terhadap Data Twitter', *Jurnal Muara*, 2(1), pp. 129–138.