p-ISSN 2477-3506
e-ISSN 2549-1938

ANALISIS SENTIMEN PADA TWEET TERKAIT VAKSIN COVID-19 MENGGUNAKAN METODE SUPPORT VECTOR MACHINE

Hashri Hayati[1], Muhammad Riza Alifi[2]

[1,2]Politeknik Negeri Bandung
Email: [1]hashri.hayati@polban.ac.id, [2]muhammad.riza@polban.ac.id

Abstrak

Abstrak-- Covid-19 adalah penyakit yang ditetapkan sebagai pandemi global sejak maret 2020. Salah satu tantangan dalam menghadapi pandemi Covid-19 saat ini adalah maraknya keraguan penggunaan vaksin, padahal vaksinasi adalah salah satu cara tersukses untuk mengatasi wabah penyakit menular. *Vaccine hesitancy* ini diantaranya dapat diamati dari sentimen atau persepsi publik di media sosial, salah satunya adalah Twitter. Keberadaan media sosial dapat mempengaruhi serapan informasi yang diterima seseorang, dalam kasus ini media sosial juga menjadi media propaganda anti vaksin yang dapat berakibat pada menurunnya kepercayaan masyarakat terhadap vaksin Covid-19. Penelitian ini bertujuan mengembangkan model klasifikasi dengan menggunakan metode *Support Vector Machine* (SVM) untuk analisis sentimen pada Tweet terkait vaksin Covid-19. Beberapa penelitian sebelumnya telah melakukan analisis sentimen terkait Covid-19, namun penelitian ini secara spesifik melakukan analisis sentimen pada topik vaksin Covid-19 sehingga persiapan data dan konfigurasi model akan berbeda. Penelitian ini juga menggunakan metodologi Design Science Research Methodology (DSRM) untuk penelitian secara keseluruhan sebelu fokus pada penggunaan metode SVM. Hasil penelitian terdiri dari algoritma pembuatan data set dan model klasifikasi untuk analisis sentimen yang dapat digunakan untuk mengetahui persepsi publik terhadap isu vaksinasi Covid-19. Penelitian ini juga membandingkan penggunaan tokenisasi unigram dan bigram. Berdasarkan hasil yang diperoleh, nilai rata-rata setiap aspek pengukuran evaluasi lebih tinggi diperoleh saat tokenisasi bigram ikut digunakan. Meskipun lebih tinggi, nilai yang diperoleh memiliki selisih yang tidak signifikan yaitu pada kisaran 0,6% - 0,7%. Pada hasil evaluasi dengan penggunaan tokenisasi unigram dan bigram, nilai tertinggi seluruh aspek pengukuran yaitu *accuracy*, *recall*, f-*measure*, dan *precision* adalah 84%.

Kata Kunci: *Analisis sentimen, SVM, Vaksin covid-19*

Abstract

Abstract-- Covid-19 is a disease that has been declared a global pandemic since March 2020. One of the challenges in dealing with the current Covid-19 pandemic is the widespread doubts about the use of vaccines, even though vaccination is one of the most successful ways to deal with infectious disease outbreaks. Vaccine hesitancy can be observed, among others, from public sentiment or perception on social media, one of them is Twitter. The existence of social media can affect the absorption of information received by a person, in this case social media is also a medium for anti-vaccine propaganda which can result in a decrease in public confidence in the Covid-19 vaccine. This study aims to develop a classification model using the Support Vector Machine (SVM) method for sentiment analysis of Tweet related to the Covid-19 vaccine. Several previous studies have conducted sentiment analysis related to Covid-19, but this research specifically conducts sentiment analysis on the topic of the Covid-19 vaccine so that data preparation and model configuration will be different. This study also uses the Design Science Research Methodology (DSRM) for research as a whole before focusing on the use of the SVM method. The results of the study consist of an algorithm for creating data sets and a

classification model for sentiment analysis that can be used to determine public perceptions of the issue of Covid-19 vaccination. This study also compares the use of unigram and bigram tokenization. Based on the results obtained, the average value of each aspect of the evaluation measurement is higher when the bigram tokenization is used. Although higher, the value obtained has an insignificant difference in the range of 0.6% - 0.7%. In the evaluation results using unigram and bigram tokenization, the highest scores for all aspects of measurement, namely accuracy, recall, f-measure, and precision were 84%.

Keywords: *Sentiment analysis, SVM, Covid-19 vaccine*

I. PENDAHULUAN

Pada akhir tahun 2019, World Health Organization (WHO) mengumumkan adanya klaster pneumonia di Wuhan, China yang kemudian dinamakan covid-19 (Boon-Itt dan Skunkan, 2020). Covid-19 menyebar dengan sangat cepat dan menginfeksi orang di berbagai belahan dunia sehingga pada bulan maret 2020 WHO mengumumkan status covid-19 sebagai pandemi global (WHO, 2020). Satu tahun setelah pengumumannya sebagai pandemi, hingga saat ini dunia belum bisa pulih dari covid-19.

Vaksinasi merupakan cara tersukses dalam intervensi kesehatan masyarakat yang dapat dilakukan untuk menghentikan wabah penyakit menular. Namun sayangnya keraguan terhadap penggunaan vaksin terus berkembang. Di tahun 2019, WHO mengumumkan *vaccine hesitancy* sebagai satu dari 10 ancaman terbesar dalam kesehatan global (Puri dkk., 2020).

Konten terkait dengan vaksin banyak tersebar di media sosial (Puri dkk., 2020). Dengan adanya pembatasan pergerakan dan perintah untuk tinggal di rumah akibat pandemik covid-19, media sosial seperti Twitter telah menjadi tempat bagi penggunanya untuk mengekspresikan kekhawatiran, opini, dan perasaan mereka terhadap covid-19. Individu, Lembaga Kesehatan, dan pemerintah juga menggunakan Twitter untuk berkomunikasi tentang covid-19 (Chandrasekaran dkk., 2020). Mengamati pembicaraan publik di Twitter terkait dengan layanan kesehatan dan kebijakan pemerintah dapat menjadi salah satu tolak ukur untuk melihat sentimen, khususnya untuk isu terkini seperti covid-19. Twitter telah banyak digunakan untuk *early warning notifier*, kanal komunikasi darurat, pemantauan persepsi publik dan sumber data pengawasan kesehatan masyarakat dalam berbagai bencana alam dan wabah penyakit (Ordun dkk., 2020).

Keberadaan media sosial dapat mempengaruhi serapan informasi yang diterima seseorang, dalam kasus ini media sosial menjadi media propaganda anti vaksin yang dapat berakibat pada menurunnya kepercayaan masyarakat terhadap vaksin covid-19 (Puri dkk., 2020). Oleh karena itu, sentimen di media sosial penting untuk diamati agar pemerintah dan lembaga kesehatan dapat menentukan strategi yang tepat dalam menangani covid-19 khususnya dalam hal komunikasi publik.

Penelitian dari Sakun Boon-Itt (Boon-Itt dan Skunkan, 2020) membahas terkait analisis sentimen dalam topik Covid-19. Penelitian tersebut menggunakan pendekatan berdasarkan *lexicon* untuk melakukan analisis sentimen. Pendekatan analisis sentimen berbasis *lexicon* memanfaatkan *lexicon* sentimen untuk menentukan polaritas konten tekstual yang diberikan. Sebuah *lexicon* atau kamus mewakili daftar kata dengan polaritas sentimen terkait (Nasim dkk., 2017). Pendekatan lain yang dapat digunakan dalam analisis sentimen adalah pendekatan *machine learning*. Pendekatan analisis sentimen berbasis *machine learning* mempelajari model prediktif menggunakan *dataset* pelatihan yang disediakan dan mengevaluasi kinerja model yang dipelajari pada *dataset* pengujian (Nasim dkk., 2017). Penelitian dari Kolchyna (Kolchyna dkk., 2015) membandingkan dua pendekatan tersebut dan hasilnya menunjukkan bahwa metode *machine learning* berbasis Support Vector Machine (SVM) dan Naive Bayes *classifier* mengungguli metode *lexicon* pada kasus analisis sentimen Tweet terkait perusahaan dari sektor ritel untuk memprediksi pergerakan harga saham. Pada pendekaan *lexicon*, penentuan sentimen didasarkan pada keberadaan kata yang menggambarkan sentimen tertentu tanpa adanya pertimbangan terhadap domain yang dibahas secara keseluruhan dalam suatu teks.

Pada penelitian dari Sakun Boon-Itt (Boon-Itt dan Skunkan, 2020) analisis sentimen dilakukan terhadap data Tweet covid-19 keseluruhan, sementara penelitian ini akan fokus pada isu vaksinasi covid-19. Penelitian terkait dengan Covid-19 saat ini dipermudah dengan terbukanya akses

terhadap banyak *dataset*. *Dataset* yang akan digunakan dalam penelitian ini adalah *dataset* yang berisi Tweet. Tweet adalah pesan dengan karakter terbatas yang diunggah seseorang dalam media sosial Twitter. Salah satu dataset Tweet terkait dengan Covid-19 yang dapat diakses secara terbuka adalah Covid-19 Tweets Dataset yang dikeluarkan Christian E. Lopez, dkk (Lopez dkk., 2020). *Dataset* ini berisi Tweets terkait Covid-19 dalam berbagai bahasa dari mulai Januari 2020 hingga Juni 2021 dan masih terus diperbarui. Penelitian ini akan menggunakan sebagian data dari *dataset* tersebut. Penelitian ini akan fokus pada isu vaksinasi covid-19 saja, sehingga perlu dilakukan pemilahan data berdasarkan lingkup bahasan Tweet yang ditetapkan.

Analisis sentimen adalah proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini (Buntoro, 2016). Pada penelitian ini analisis sentimen dilakukan dengan pendekatan *machine learning*, khususnya dengan menggunakan *supervised learning*. Terdapat beberapa tahapan dalam proses analisis sentimen, yaitu (Hadna dkk., 2016) pengumpulan dan pelabelan data, *pre-processing*, pembobotan kata, analisis sentimen, pengukuran kualitas hasil uji.

Terdapat beberapa metode yang dapat dilakukan untuk analisis sentimen (Abirami dan Gayathri, 2017). Penelitian ini akan menggunakan SVM untuk analisis sentimen, yaitu pembuatan model klasifikasi untuk mengelompokan Tweet ke dalam tiga kelas yaitu positif, negatif, netral. Metode SVM dipilih karena menurut penelitian sebelumnya (Abirami dan Gayathri, 2017; Hadna dkk., 2016) SVM merupakan salah satu metode yang dapat digunakan untuk analisis sentimen dan cenderung memberikan hasil lebih baik dibandingkan Naïve Bayes.

Beberapa penelitian terkait dengan analisis sentimen juga menggunakan SVM, diantaranya:

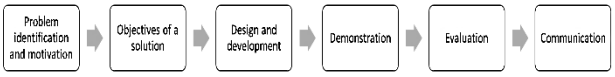
a. Analisis Sentimen Hatespeech Pada Twitter Dengan Metode *Naïve Bayes Classifier* Dan *Support Vector Machine* (Buntoro, 2016)
b. Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Twitter *Dengan Metode Support Vector Machine* dan *Lexicon Based Features* (Rofiqoh dkk., 2017)
c. Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan *Metode Support Vector Machine* (Novantirani dkk., 2015)
d. Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter (Buntoro 2017)
e. Analisis Sentimen Twitter untuk Teks Berbahasa Indonesia dengan *Maximum Entropy* dan *Support Vector Machine* (Putranti dan Winarko, 2014)

Berdasarkan penelitian-penelitian tersebut dapat disimpulkan bahwa SVM dapat digunakan untuk melakukan analisis sentimen, khususnya untuk topik-topik yang ramai diperbincangkan publik. Penelitian yang akan dilakukan adalah analisis sentimen pada tweet terkait vaksin Covid-19 dengan metode SVM. Hasil penelitian ini adalah model klasifikasi untuk analisis sentimen yang dapat digunakan untuk mengetahui persepsi publik terhadap isu vaksin covid-19.

Fokus kajian pada penelitian ini adalah penerapan teknologi. Penerapan teknologi *machine learning* khususnya SVM sudah banyak digunakan dalam kasus sentimen analisis namun penerapan tersebut tetap memerlukan kajian akademis karena sangat bergantung dengan studi kasus yang dipilih. Meskipun penelitian ini menggunakan metode yang sudah ada, yaitu SVM, kajian pada penelitian tetap dibutuhkan terutama terkait penentuan konfigurasi yang tepat agar dihasilkan model klasifikasi untuk analisis sentimen yang baik. Konfigurasi yang dimaksud adalah penentuan parameter untuk menyeleksi data yang akan digunakan seperti kata kunci, pemilihan teknik pra-pemrosesan, parameternya, dan data terkait seperti kamus istilah di domain studi kasus, parameter saat pembangunan model, dan lain-lain.

II. METODE

Penelitian ini mengikuti *Design Science Research Methodology* (DSRM) sebagai metodologi penelitian (Peffers dkk., 2007). Proses utama yang dijalankan pada penelitian dengan metode DSRM digambarkan pada Gambar 1.



Gambar 1. *Design Science Research Methodology*

2.1. Identifikasi Masalah

Pada penelitian ini identifikasi masalah diawali dengan melakukan studi pustaka terkait penelitian sejenis yang mencakup topik penerapan metode klasifikasi untuk analisis sentimen, covid Tweet *dataset*, dan evaluasi model klasifikasi. Studi pustaka bertujuan untuk merumuskan masalah yang akan dikaji dalam penelitian ini, yaitu bagaimana

analisis sentimen dilakukan pada Tweet terkait vaksin Covid-19.

2.2. Tujuan Solusi

Sebagaimana telah dijelaskan pada bagian pendahuluan, telah ada beberapa penelitian terkait baik dari sisi metode klasifikasi yang dipilih maupun kajian studi kasus yaitu Tweet tentang Covid-19. Namun penelitian ini akan fokus pada lingkup topik vaksin saja. Lingkup topik pembicaraan di Twitter berimplikasi pada pemilahan kata kunci dan konfigurasi lain, serta cara pembuatan data set yang mungkin berbeda. Oleh karena itu tujuan dari solusi yang dihasilkan pada penelitian ini adalah menghasilkan model klasifikasi untuk analisis sentimen Tweet terkait vaksin Covid-19. Evaluasi dari model klasifikasi sangat bergantung pada data dan pemrosesan yang dilakukan, sehingga hasil evaluasi dengan data yang berbeda tidak dapat secara langsung dibandingkan. Pada penelitian (Hadna dkk., 2016) dilakukan komparasi beberapa metode klasifikasi pada beberapa studi kasus. Berdasarkan data pada penelitian tersebut diperoleh nilai akurasi klasifikasi menggunakan SVM ada di rentang 60,05% - 96,76%. Penelitian ini mengharapkan hasil evaluasi yang setidaknya berada di nilai tengah rentang tersebut, yaitu sekitar 80%.

2.3. Perancangan dan Pengembangan

Hasil dari perancangan pada penelitian ini terdiri dari:

a. Algoritma pembuatan *dataset*

Terdapat dua jenis sumber *dataset* yang biasa digunakan, yaitu dengan menggunakan *dataset* yang telah tersedia dan mengambil data dari sumber data secara langsung. Penelitian ini menggunakan data yang berasal dari *dataset* yang telah ada, yaitu *dataset* yang disediakan oleh penelitian (Lopez dkk., 2020). Namun untuk menggunakan *dataset* tersebut dibutuhkan proses tersendiri, karena data yang tersedia berupa Tweet ID, sementara pemilahan data berdasarkan kata kunci perlu dilakukan terhadap konten Tweet. Oleh karena itu salah satu luaran dari proses perancangan dan pengembangan adalah algoritma pembuatan *dataset* untuk menghasilkan *dataset* yang telah sesuai dengan cakupan topik yang diinginkan, yaitu Tweet terkait vaksinasi covid-19.

b. Arsitektur pengembangan model klasifikasi

Arsitektur pengembangan model klasifikasi berisi tahapan dan konfigurasi yang dibutuhkan untuk membangun model. Informasi ini diperlukan agar proses pengembangan dapat dipertanggungjawabkan dan dikembangkan lebih lanjut. Penelitian ini akan menghasilkan arsitektur pengembangan dari dua sisi, yaitu metode atau pemrosesan yang dilakukan, serta *tools* atau *library* yang digunakan untuk membangunnya. Arsitektur pengembangan model akan mencakup *pre-processing* data dan pembuatan model. *Pre-processing* dilakukan terhadap *dataset* yang telah dihasilkan di tahapan sebelumnya. *Pre-processing* dilakukan agar model yang dihasilkan dapat lebih baik performanya. Beberapa *pre-processing* yang akan dilakukan pada penelitian ini adalah normalisasi kata, anotasi, *unpack hash tag, unpack contraction*, dan *transform emoticon.* Ekstraksi fitur kemudian dilakukan terhadap data yang telah di *pre-processing*. Pembuatan model klasifikasi dilakukan dengan *supervised learning* menggunakan metode SVM. Model klasifikasi dibangun dengan memasukan *dataset* yang telah diberi label sentimen (positif, negatif, netral).

SVM secara resmi diperkenalkan oleh Cortes dan Vapnik pada tahun 1995 (Cortes dan Vapnik, 1995) dan terbukti menjadi salah satu algoritma *supervised machine learning* yang banyak digunakan untuk tujuan klasifikasi (Ahmad dkk., 2017). Metode SVM membangun satu atau sekumpulan *hyper-plane* dalam ruang berdimensi tinggi atau tak terbatas, yang dapat digunakan untuk klasifikasi, regresi atau tugas lainnya. Secara intuitif, pemisahan yang baik dicapai oleh *hyper-plane* yang memiliki jarak terbesar ke titik data latih terdekat dari kelas mana pun (disebut *margin* fungsional), karena secara umum semakin besar *margin*, semakin rendah kesalahan generalisasi pengklasifikasi (Pedregosa dkk., 2011).

c. Rancangan alur pengujian model klasifikasi

Rancangan alur pengujian akan dihasilkan pada tahapan ini untuk menjelaskan rencana evaluasi yang akan dilakukan terhadap model yang dihasilkan. Pada tahapan ini akan dikaji parameter apa yang akan diukur serta bagaimana proses pengukuran atau metode yang digunakan.

2.4. Demonstrasi

Pada aktivitas ini, rancangan yang dihasilkan dari tahap sebelumnya akan diimplementasikan. Penelitian ini akan mendemonstrasikan hasil rancangan dengan mengembangkan model klasifikasi menggunakan bahasa pemrograman Python dan beberapa *tools* yang dibutuhkan seperti Google Collaboratory, Google Drive, dan beberapa *library*.

2.5. Evaluasi

Rancangan alur pengujian model klasifikasi yang dihasilkan di tahap 2.3 akan dilaksanakan pada aktivitas evaluasi. Penelitian ini akan mengevaluasi model dengan melakukan validasi dan pengukuran kinerja model. Model yang telah dihasilkan divalidasi menggunakan *10-fold cross validation*,

yaitu membagi jumlah data ke dalam 10 kelompok dimana secara bergilir 1/10 bagian dijadikan sebagai data uji dan sisanya sebagai data latih sebanyak 10 kali. Evaluasi kinerja model dilakukan dengan menghitung nilai *accuracy, precission, recall,* dan f-*measure*. Pengukuran dilakukan berdasarkan *confusion matrix* seperti pada Tabel I. Formula 1, 2, 3, dan 4 menunjukkan formula yang digunakan untuk setiap nilai evaluasi yang digunakan (Awad dan Khanna, 2015).

2.6. Komunikasi

Komunikasi hasil penelitian dilakukan dengan publikasi artikel ilmiah ini.

Tabel 1. *Confusion Matrix*

		Actual Class	
		Positive	*Negative*
Predicted Class	*Positive*	*True Positive* (TP)	*False Positive* (FP)
	Negative	*False Negative* (FN)	*True Negative* (TN)

$$accuracy\ (\mathrm{AC}) = \frac{TP+TN}{TP+TN+FN+FP} \quad (1)$$

$$precision\ (\mathrm{P}) = \frac{TP}{TP+FP} \quad (2)$$

$$recall\ (\mathrm{R}) = \frac{TP}{TP+FN} \quad (3)$$

$$f-measure = \frac{(\beta^2+1)\times P\times R}{(\beta^2\times P)+R} \quad (4)$$

III. HASIL DAN PEMBAHASAN

Pada bagian ini akan dijelaskan hasil dan pembahasan penelitian yang terbagi dalam 3 bagian, yaitu pembuatan *dataset*, pengembangan model klasifikasi, dan evaluasi model klasifikasi.

3.1. Pembuatan Data Set

Proses pembuatan *dataset* secara umum terdiri dari tiga tahapan, yaitu menentukan kata kunci untuk pencarian data, menentukan sumber data, dan mengumpulkan data sesuai kata kunci. Penentuan kata kunci dibutuhkan untuk memilah data yang akan digunakan agar sesuai dengan lingkup studi kasus. Penentuan sumber data dilakukan untuk menentukan cara pengumpulan data. Tahap terakhir adalah pengumpulan data yang disesuaikan dengan kata kunci dan sumber data yang telah ditentukan. *Dataset* yang digunakan pada penelitian ini dibatasi hanya untuk Tweet berbahasa Inggris.

3.1.1. Penentuan Kata Kunci

Penentuan kata kunci pada penelitian ini dilakukan dengan mengamati pembicaraan di Twitter terkait vaksin Covid-19. Terdapat dua kata kunci utama yang menjadi lingkup kasus dalam penelitian ini, yaitu ‘covid-19’ dan ‘vaksin’. Dua kata kunci tersebut diturunkan menjadi beberapa kata kunci yang merepresentasikan sinonim atau kata lain yang berkaitan seperti nama merek vaksin dan lain-lain.

Pada penelitian ini kata kunci terkait ‘covid-19’ yang digunakan adalah sinonim atau kata yang biasa digunakan untuk menggantikan kata ‘covid-19’ di Twitter seperti ‘coronavirus’, ‘ncov19’, dan lain-lain, serta kata yang sering digunakan saat pengguna Twitter membuat Tweet terkait covid-19 seperti ‘stayathome’. Penelitian ini menggunakan kata kunci yang digunakan pada penelitian (Lopez dan Gallemore, 2020) untuk kata kunci yang terkait atau merepresentasikan topik covid-19.

Kata kunci terkait ‘vaksin’ pada penelitian ini terdiri dari tiga kategori, yaitu sinonim dari kata ‘vaksin’, merek dari beberapa vaksin, dan organisai pembuat vaksin. Sinonim kata ‘vaksin’ yang dijadikan kata kunci pada penelitian ini adalah ‘*vaccine’* dan ‘*immunization’*. Merek dan organisasi pembuat vaksin saat ini berjumlah sangat banyak. Pada penelitian ini, merek dan organisasi pembuat vaksin yang dijadikan kunci adalah vaksin yang terdaftar telah lulus uji klinis tahap 3 (sedang ada di tahap 4) berdasarkan data *vaccine tracker* WHO versi 18 Juni 2021 (WHO, 2021). Berdasarkan data tersebut terdapat 4 merek vaksin yang memenuhi kriteria. Dari empat vaksin tersebut diturunkan menjadi 13 kata kunci sebagaimana tertulis pada anggota himpunan V di bawah.

Berdasarkan analisis kata kunci tersebut, maka data yang akan digunakan dalam penelitian ini adalah Tweet yang mengandung setidaknya satu kata kunci terkait ‘covid-19’ dan setidaknya satu kata kunci terkait ‘vaksin’ seperti diformulasikan pada Formula (5). Artinya, data yang dapat digunakan adalah data yang mengandung kata kunci (‘coronavirus’, ‘*vaccine’*), (‘covid’, ‘sinovac), (‘*stayathome’*, ‘Moderna’), dan kombinasi lain.

C = {‘virus’, ‘coronavirus’, ‘ncov19’, ‘ncov2019’, ‘covid’, ‘rona’, ‘ramadandirumah’, ‘dirumahaja’, ‘*stayathome’*}

V = {‘*vaccine’*, ‘*immunization’*, ‘coronavac’, ‘sinovac’, ‘ChAdOx1’, ‘AZD1222’, ‘AstraZeneca’, ‘mRNA-1273’, ‘CanSino’, ‘Moderna’, ‘BNT162b2’, ‘Comirnaty’, ‘Pfizer’, ‘BioNTech’, ‘Fosun Pharma’}

$$K = \{ (x,y) \mid x \in C, y \in V \} \quad (5)$$

C = Kata kunci terkait covid-19 (Lopez dan Gallemore, 2020)

V = Kata kunci terkait vaksin

K = Kata kunci yang harus ada dalam data Tweet yang akan digunakan

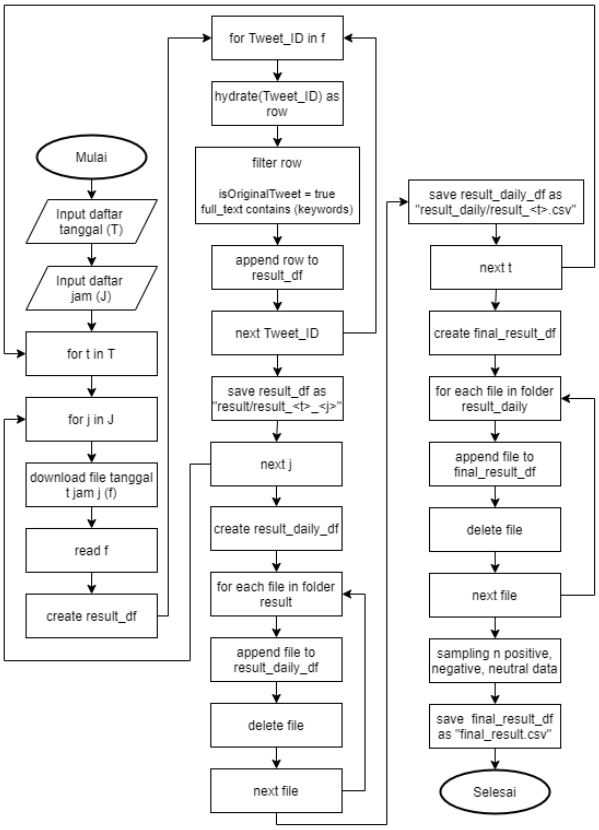
3.1.2. Pemilihan Sumber Data

Untuk membuat model klasifikasi dengan pendekatan *machine learning*, dibutuhkan data set yang berisi data yang akan diklasifikasi dan label kelasnya. Penelitian ini menggunakan data Tweet terkait vaksin Covid-19. Label setiap data Tweet dibutuhkan pada *supervised learning* sebagai referensi bagi model dalam mengklasifikasikan data. Pada penelitian ini data awal diambil dari penelitian (Lopez dan Gallemore, 2020) yang dapat diakses melalui Github. Penggunaan data set ini dipilih karena data set ini telah menyeleksi Tweet sesuai dengan kasus yang sesuai dengan penelitian yaitu Covid-19. Selain itu data set ini juga memungkinkan penggunaan data melebihi batas waktu yang ditetapkan oleh API Twitter jika mengambil data langsung. Data set ini juga dipilih karena telah tersedianya label sentimen yang dapat digunakan untuk referensi awal pembuatan model, sehingga proses pelabelan tidak perlu dilakukan pada penelitian ini.

3.1.3. Pengumpulan Data

Proses pengumpulan data sangat tergantung dengan sumber data yang dipilih. Data set yang digunakan pada penelitian ini adalah data set yang mengandung label sentimen, dimana data set diorganisasi per jam. Penelitian ini menggunakan data dalam rentang waktu tiga bulan terakhir, yaitu bulan Maret 2021 sampai Mei 2021. Jumlah data yang tersedia di data set ini sekitar 170 ribu Tweet per hari atau sekitar 35 juta Tweet per bulan. Data tersebut merupakan data yang sudah terseleksi dengan kata kunci terkait ‘covid’ (himpunan C pada formula (5)). Data yang digunakan adalah Summary_SentimentTable yang memiliki atribut Tweet_ID, label sentimen, dan *non-normalized prediction* untuk setiap kelas sentimen (Lopez dan Gallemore, 2020).

Data yang diperoleh dari sumber data perlu diolah untuk dijadikan *dataset* penelitian. Pengolahan dilakukan dengan menyeleksi Tweet yang mengandung kata kunci terkait ‘vaksin’ (himpunan V pada formula (5)) dan Tweet original. Penyeleksian data tidak dapat langsung dilakukan terhadap data set karena data set hanya menampilkan Tweet_ID tanpa konten dari Tweet tersebut. Oleh karena itu perlu dilakukan *hydration* terhadap Tweet_ID tersebut. Proses pengumpulan data mengalami tantangan yaitu besarnya jumlah Tweet yang harus di-*hydration* untuk diseleksi tidak dapat dilakukan sekaligus karena terbatasnya sumber daya penelitian. Setiap *file* di sumber data yang berisi data mentah Tweet setiap jam yang berhasil dikumpulkan kebanyakan memiliki ukuran *file* dalam rentang 5MB – 10 MB. Untuk mengambil data satu hari saja dibutuhkan sekitar 200MB. Ukuran tersebut hanya untuk mengunduh data mentah sebelum di-*hydrate*. Hasil *hydration* membutuhkan penyimpanan yang jauh lebih besar karena memuat lebih banyak data. Oleh karena itu, penelitian ini menghasilkan algoritma pengambilan data yang dapat dilakukan dengan fasilitas penyimpanan terbatas dan waktu *running* aplikasi terbatas. Karena jumlah data dengan label positif, negatif, dan netral tidak seimbang, data yang berhasil diseleksi kembali di-*sampling* sejumlah data milik kelas dengan data paling sedikit. Algoritma pembuatan data set yang digunakan pada penelitian ini digambarkan pada Gambar 2.



Gambar 2. Algoritma pembuatan *dataset*

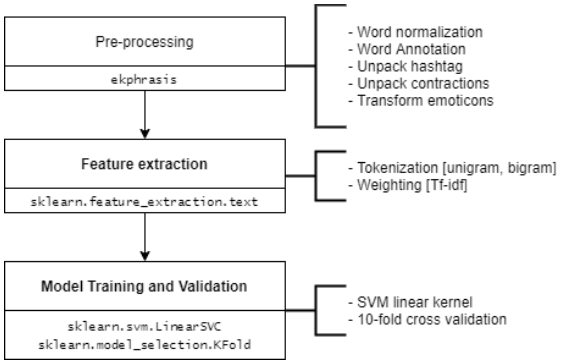
Pada penelitian ini rentang tanggal yang digunakan adalah 1 Maret 2021 sampai 11 Mei 2021, menyesuaikan dengan ketersediaan data saat penelitian berlangsung. Karena sumber daya yang terbatas, penelitian ini hanya menggunakan sebagian data yaitu data pada tanggal ganjil dan satu per tiga data yang tersedia dalam satu hari yaitu data per jam pada jam 00, 03, 06, 09, 12, 15, 18 dan 21. Dari hasil data yang terkumpul, data dengan label positif secara signifikan berjumlah lebih sedikit dari data dua kelas lainnya. Oleh karena itu penelitian melakukan *sampling* dengan hasil akhir jumlah data

yang digunakan dalam data set sebanyak 360000 data, dengan masing-masing 120000 data berlabel positif, negatif, dan netral. Data hasil *hydration* memiliki banyak atribut, namun penelitian ini hanya menggunakan atribut '*full_text*' dari data Tweet yang dipilih.

3.2. Pengembangan Model Klasifikasi Analisis Sentimen

Pengembangan model klasifikasi pada analisis sentimen terdiri dari beberapa tahapan yaitu *pre-processing* data, *feature extraction*, *model training and validation*. Tweets biasanya terdiri dari kalimat yang tidak lengkap, kurang terstruktur dan banyak *noise*, ekspresi yang tidak beraturan, kata yang tidak sempurna bentuknya, dan istilah yang tidak ada di kamus (Jianqiang dan Xiaolin, 2017). Oleh karena itu perlu dilakukan *pre-processing* terhadap data untuk mengurangi *noise*. Untuk membuat model klasifikasi, data harus berbentuk angka. Tahap *feature extraction* dilakukan untuk mentransformasi teks menjadi angka. Setelah data ditransformasi, akan dilakukan proses *training* untuk menghasilkan model klasifikasi dan dilakukan validasi. Algoritma yang digunakan untuk *training* model adalah SVM (Support Vector Machine). Validasi terhadap model dilakukan dengan 10-*fold cross validation*.

Penelitian ini menggunakan Google Collaboratory dan Google Drive untuk alat pengembangan, bahasa pemrograman python dan beberapa *library* dari scikit learn. Gambar 3 menunjukan tahapan, *library* yang digunakan, dan teknik atau parameter yang diterapkan dalam pengembangan model klasifikasi sentimen di penelitian ini.



Gambar 3. Algoritma pengembangan model klasifikasi analisis sentimen

3.2.1. Data *Pre-processing*

Pada penelitian ini *pre-processing* dilakukan dengan menggunakan Notebook di Google Collaboratory yang berbasis python dan menggunakan library ekphrasis (Baziotis, 2017). Terdapat beberapa metode *pre-processing* data yang digunakan dalam penelitian ini, yaitu:

a. Normalisasi kata

Normalisasi kata dilakukan untuk menghilangkan atau mengubah menjadi kata yang sama beberapa jenis kata yang semestinya tidak mempengaruhi sentimen di suatu kalimat. Penelitian ini melakukan normalisasi terhadap URL, *e-mail*, *user*, dan tanggal.

b. Anotasi kata

Dalam teks non-formal seperti Twitter, pengguna seringkali menggunakan kata dalam bentuk tidak baku untuk menekankan hal tertentu. Contohnya penggunaan *caps lock*, *hash tag*, kata dengan huruf yang dipanjangkan, atau kata yang diulang-ulang. Meskipun memiliki makna kata yang sama, cara penulisan dapat menunjukan sentimen atau emosi dari pengguna. Penelitian ini melakukan anotasi dan pengubahan kata ke bentuk dasar untuk menjaga informasi penekanan yang ditulis pengguna dengan tetap menggunakan kata dasar sesuai kamus.

c. Unpack hashtag

Penggunaan *hashtag* merupakan salah satu ciri khas teks di Twitter atau media sosial secara umum. *Hashtag* dapat terdiri dari satu atau beberapa kata. Penggunaan beberapa kata dalam satu *hashtag* tidak dipisahkan dengan spasi sebagaimana pada kalimat biasa. Oleh karena ini penelitian ini juga melakukan *unpack hashtag*, yaitu memisahkan *hashtag* dengan spasi untuk setiap kata. Pemisahan kata didasarkan pada penggunaan *camelCase*, misalnya #stayAtHome akan diubah menjadi # stay At Home.

d. *Unpack contraction*

Terdapat beberapa kata umum yang sering disingkat dalam penulisannya seperti *can't, wouldn't*, dan sebagainya. *Unpack contraction* akan mengubah kata yang disingkat tersebut menjadi bentuk dasarnya, contohnya *can't* dipanjangkan menjadi *can not*.

e. *Transform emoticon*

Emoticon terkadang digunakan dalam teks Twitter. *Emoticon* memiliki makna yang dapat mempengaruhi sentimen, namun karena tidak terdaftar dalam kamus atau *corpus*, hal ini sering terlewatkan. Pada penelitian ini dilakukan transformasi dari *emoticon* menjadi kata yang mewakili emosi pada *emoticon* tersebut seperti *emoticon* :) diubah menjadi kata *smile*.

3.2.2. *Feature Extraction*

Algoritma *machine learning* seperti SVM membutuhkan data dalam format angka untuk menghasilkan model klasifikasi. Data yang dimiliki adalah data teks, oleh karena itu ekstraksi fitur dilakukan untuk mentransformasi teks menjadi angka dalam format yang dapat diterima algoritma. Terdapat beberapa metode yang dapat dilakukan untuk mengekstraksi fitur. Penelitian ini

menggunakan metode bigram dan unigram serta Tf-Idf untuk melakukan ekstraksi fitur.

Unigram merupakan metode tokenisasi yang memisahkan kalimat menjadi token yang terdiri dari 1 kata. Contohnya kalimat "*I get vaccine number 2!*" akan ditokenisasi menjadi ["*I*", "*get*", "*vaccine*", "*number*", "*2*!"]. Namun terkadang penggunaan tokenisasi unigram dapat menghilangkan makna dari frase yang terdiri dari dua kata seperti "New York" yang memiliki makna berbeda dengan komponen katanya yaitu "New" dan "York". Oleh karena itu, penelitian ini juga menambahkan tokenisasi bigram, sehingga selain token yang terdiri dari 2 kata. Pada penggunaan unigram dan bigram, contohnya kalimat "*I get vaccine number 2!*" akan ditokenisasi menjadi ["*I*", "*get*", "*vaccine*", "*number*", "*2!*", "*I get*", "*get vaccine*", "*vaccine number*", "*number 2*!"]. Penelitian ini menggunakan CountVectorizer dari sub modul sklearn.feature_extraction.text library Scikit-learn (Pedregosa dkk., 2011) untuk mengubah kumpulan dokumen teks menjadi matriks jumlah token. Parameter yang digunakan pada penelitian ini adalah ngram_range = (1,2) yang artinya tokenisasi dilakukan dengan 1-gram dan 2-gram. Karena matriks yang dihasilkan bersifat *sparse* atau tersebar dengan banyak nilai 0, penelitian ini juga menggunakan modul spicy.sparse untuk merepresentasikannya.

Penelitian ini selanjutnya melakukan pembobotan kata untuk membedakan kata yang kontributif terhadap penentuan kelas. Dalam pembuatan kelas terdapat beberapa kata yang tinggi kemunculannya di setiap kalimat namun tidak memiliki makna yang signifikan, contohnya kata "*the*", "*is*", "*are*", dan lain-lain. Kata yang tidak memberi makna yang signifikan atau tidak mempengaruhi penentuan kelas dari data tersebut seharusnya memiliki bobot yang lebih kecil atau bahkan tidak diperhitungkan. Penelitian ini menggunakan metode Tf-idf untuk melakukan pembobotan kata agar diperoleh hasil yang lebih optimal. Tf-idf (*Term Frequency-Inverse Document Frequency*) adalah pembobotan yang menggabungkan konsep frekuensi kemunculan suatu *term* atau kata dalam suatu dokumen atau kalimat (*term frequency)* dan jumlah dokumen yang memiliki kemunculan *term* tertentu (*document frequency*) (Hadna dkk., 2016). Tf-idf mengukur seberapa penting kata tertentu terhadap dokumen dan seluruh korpus. Kata-kata yang langka dalam dokumen akan memiliki skor tinggi dalam Tf-idf. Penelitian ini menggunakan TfidfTransformer dari submodul sklearn.feature_extraction.text library Scikit-learn (Pedregosa dkk., 2011) untuk mengubah matriks hasil vektorisasi dari token menjadi representasi Tf-idf.

3.2.3. Model Training and Validation

Model klasifikasi dibangun dengan melakukan *training* menggunakan algoritma *machine learning*. Algoritma yang digunakan pada penelitian ini adalah SVM, sementara data latih yang digunakan adalah matriks hasil tokenisasi unigram dan bigram dan pembobotan kata dengan Tf-Idf yang dihasilkan di tahap sebelumnya. Penelitian ini menggunakan sub modul LinearSVC dari modul SVM library Scikit-learn (Pedregosa dkk., 2011). Berdasarkan hasil pengujian untuk pencarian parameter yang menghasilkan akurasi maksimal, penelitian ini menggunakan daftar parameter yang dituliskan di Tabel 2.

Tabel 2. Parameter model

Parameter	Nilai
kernel	linear
penalty	l2
loss	squared_hinge
multi class	ovr

Pembuatan dan validasi model dilakukan dengan metode 10-*fold cross validation*, yaitu metode yang membagi seluruh data menjadi 10 potongan dimana 1/10 bagian dijadikan data tes dan sisanya menjadi data latih. Penelitian ini menggunakan total 360000 data, sehingga dilakukan 10 kali pembuatan model dengan proporsi 90% data latih (324000) dan 10% data uji (36000) secara bergiliran.

3.3. Evaluasi Model Klasifikasi

Terdapat 4 nilai yang dievaluasi dalam penelitian ini, yaitu *accuracy*, *precission*, *recall*, dan f-*measure* sebagaimana dijelaskan pada subbab 2.5. Tabel III dan Tabel IV menunjukan hasil evaluasi model pada 10-*fold cross validation* yang dilakukan. Tabel III menggunakan metode tokenisasi 1-gram atau unigram, sementara Tabel IV menggunakan metode tokenisasi unigram dan bigram.

Tabel 3. Hasil evaluasi model dengan unigram

n	***precision***	***recall***	***F-measure***	***accuracy***
1	83,00%	83,33%	83,00%	83,00%
2	83,33%	83,67%	83,33%	83,00%
3	82,67%	83,00%	82,67%	83,00%
4	83,00%	82,67%	82,67%	83,00%
5	83,00%	83,00%	83,00%	83,00%
6	84,00%	83,33%	83,33%	84,00%
7	83,33%	82,67%	83,33%	83,00%
8	83,00%	83,00%	83,33%	83,00%
9	83,00%	83,00%	83,33%	83,00%
10	82,67%	82,67%	82,67%	83,00%
Average	83,10%	83,03%	83,07%	83,10%
Max	84,00%	83,67%	83,33%	84,00%
Min	82,67%	82,67%	82,67%	83,00%

Tabel 4. Hasil evaluasi model dengan unigram dan bigram

n	*precision*	*recall*	*F-measure*	*accuracy*
1	83,33%	83,67%	84,00%	84,00%
2	83,67%	84,00%	83,67%	84,00%
3	83,67%	83,33%	83,33%	83,00%
4	84,00%	83,67%	83,67%	84,00%
5	83,67%	83,67%	84,00%	84,00%
6	84,00%	84,00%	84,00%	84,00%
7	84,00%	83,67%	84,00%	84,00%
8	83,00%	83,00%	83,33%	83,00%
9	83,67%	83,33%	83,67%	83,00%
10	84,00%	84,00%	84,00%	84,00%
Average	83,70%	83,63%	83,77%	83,70%
Max	84,00%	84,00%	84,00%	84,00%
Min	83,00%	83,00%	83,33%	83,00%

Karena menggunakan 10-*fold cross validation* maka diperoleh 10 nilai hasil evaluasi dari setiap set eksperimen. Berdasarkan hasil yang diperoleh, nilai rata-rata setiap aspek pengukuran evaluasi lebih tinggi diperoleh saat tokenisasi bigram digunakan. Meskipun lebih tinggi, nilai yang diperoleh memiliki selisih yang tidak signifikan yaitu pada kisaran 0,6% - 0,7%. Hal ini dapat terjadi karena pada *dataset* dan domain yang diteliti, yaitu terkait vaksin covid-19, frase atau istilah yang terdiri dari dua kata seperti nama orang, nama tempat, dan lain-lain yang baru dapat diperoleh konteks atau maknanya saat tokenisasi bigram digunakan, tidak banyak mempengaruhi hasil klasifikasi. Pada hasil evaluasi dengan penggunaan tokenisasi unigram dan bigram, nilai tertinggi seluruh aspek pengukuran yaitu *accuracy*, *recall*, *f-measure*, dan *precision* adalah 84%.

IV. PENUTUP

Kesimpulan

Penelitian ini berhasil membangun model klasifikasi untuk analisis sentimen pada tweet terkait vaksin Covid-19 dengan metode SVM. Penelitian ini mengkaji konfigurasi yang tepat untuk menghasilkan model klasifikasi yang berkualitas. Penelitian ini menggunakan *dataset* yang sudah tersedia, namun karena jumlah data yang sangat banyak dan sumber daya penelitian yang minim, penelitian ini menghasilkan algoritma pembentukan *dataset* yang dapat mengatasi kondisi tersebut. Dari data set tersebut kemudian penelitian ini berhasil membangun model klasifikasi analisis sentimen yang dimulai dengan data *pre-processing*, *feature extraction*, dan model *training and validation*. Penelitian ini juga membandingkan penggunaan tokenisasi unigram dan bigram. Berdasarkan hasil yang diperoleh, nilai rata-rata setiap aspek pengukuran evaluasi lebih tinggi diperoleh saat tokenisasi bigram ikut digunakan. Meskipun lebih tinggi, nilai yang diperoleh memiliki selisih yang tidak signifikan yaitu pada kisaran 0,6% - 0,7%. Dengan konfigurasi parameter dan *library* yang digunakan, hasil evaluasi dengan penggunaan tokenisasi unigram dan bigram, nilai tertinggi seluruh aspek pengukuran yaitu *accuracy*, *recall*, f-*measure*, dan *precision* adalah 84%.

Saran

Hasil penelitian ini dapat dilanjutkan untuk eksploratori data dari Tweet Vaksin Covid-19. Dengan model yang dihasilkan, analisis sentimen dapat dilakukan terhadap data yang berjalan saat ini dan menghasilkan *insight* terkait vaksin covid-19 berdasarkan obrolan di media sosial khususnya Twitter. Hasil penelitian juga dapat dijadikan komponen *social media monitoring*. Kualitas model juga dapat ditingkatkan di penelitian berikutnya dengan mengkaji efek dari setiap tahapan yang dilakukan seperti membandingkan pengaruh setiap tahap *pre-processing* dan lainnya.

V. DAFTAR PUSTAKA

Abirami, A. M., & Gayathri, V. 2017. A survey on sentiment analysis methods and approach. In 2016 Eighth International Conference on Advanced Computing (ICoAC) (pp. 72-76). IEEE.

Ahmad, M., Aftab, S., & Ali, I. 2017. Sentiment analysis of tweets using SVM. International Journal of Computer Applications, 177(5), 25-29.

Awad M., Khanna R. 2015. Machine Learning. In: Efficient Learning Machines. Apress, Berkeley, CA.

Baziotis, C., Pelekis, N., & Doulkeridis, C. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017) (pp. 747-754).

Boon-Itt, S., & Skunkan, Y. 2020. Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study. JMIR Public Health and Surveillance, 6(4), e21978.

Buntoro, G. A. 2016. Analisis Sentimen Hatespeech Pada Twitter Dengan Metode Naive Bayes

Classifier Dan Support Vector Machine. Jurnal Dinamika Informatika, 5(2).

Buntoro, G. A. 2017. Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter. INTEGER: Journal of Information Technology, 2(1).

Chandrasekaran, R., Mehta, V., Valkunde, T., & Moustakas, E. 2020. Topics, Trends, and Sentiments of Tweets About the COVID-19 Pandemic: Temporal Infoveillance Study. Journal of medical Internet research, 22(10), e22624.

Cortes, C., & Vapnik, V. 1995. Support vector machine. Machine learning, 20(3), 273-297. Kluwer Academic Publishers, Boston.

Hadna, N. M. S., Santosa, P. I., & Winarno, W. W. 2016. Studi literatur tentang perbandingan metode untuk proses analisis sentimen di Twitter. Semin. Nas. Teknol. Inf. dan Komun, 2016, 57-64.

Jianqiang, Z., & Xiaolin, G. 2017. Comparison research on text pre-processing methods on twitter sentiment analysis. IEEE Access, 5, 2870-2879.

Kolchyna, O., Souza, T. T., Treleaven, P., & Aste, T. 2015. Twitter sentiment analysis: Lexicon method, machine learning method and their combination. arXiv preprint arXiv:1507.00955.

Lopez, C. E., & Gallemore, C. 2020. An Augmented Multilingual Twitter Dataset for Studying the COVID-19 Infodemic.

Lopez, C. E., Vasu, M., & Gallemore, C. 2020. Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset. arXiv preprint arXiv:2003.10359.

Nasim, Z., Rajput, Q., & Haider, S. 2017. Sentiment analysis of student feedback using machine learning and lexicon based approaches. International Conference on Research and Innovation in Information Systems (ICRIIS) 2017.

Novantirani, A., Sabariah, M. K., & Effendy, V. 2015. Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode Support Vector Machine. eProceedings of Engineering, 2(1).

Ordun, C., Purushotham, S., & Raff, E. 2020. Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs. arXiv preprint arXiv:2005.03082.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. 2007. A Design Science Research Methodology for Information Systems Research. Journal of Management Information Systems, 24(3), 45–77. doi:10.2753/mis0742-1222240302

Puri, N., Coomes, E. A., Haghbayan, H., & Gunaratne, K. 2020. Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases. Human Vaccines & Immunotherapeutics, 1-8.

Putranti, N. D., & Winarko, E. 2014. Analisis sentimen twitter untuk teks berbahasa Indonesia dengan maximum entropy dan support vector machine. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 8(1), 91-100.

Rofiqoh, U., Perdana, R. S., & Fauzi, M. A. 2017. Analisis sentimen tingkat kepuasan pengguna penyedia layanan telekomunikasi seluler indonesia pada twitter dengan metode Support Vector Machine dan Lexicon Based Features. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN, 2548, 964X.

World Health Organization, 11 Maret 2020. Diakses dari https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020, pada 23 September 2021

World Health Organization, 18 Juni 2021. Diakses dari https://www.who.int/publications/m/item/draft-landscape-of-covid-19-candidate-vaccines, pada 22 Juni 2021.