 ISSN:2302-8149
e-ISSN:2540-9719

Text Classification untuk Menganalisis Sentimen Pendapat Masyarakat Indonesia terhadap Vaksinasi Covid-19

Text Classification for Analysing Indonesian People's Opinion Sentiment for Covid-19 Vaccination

[1]Eka Miranda*, [1]Veronica Gabriella, [1]Sriyanda Afrida Wahyudi, [1]Jennifer Chai
[1]Information Systems Department, School of Information Systems, Bina Nusantara University
Jl. Raya Kb. Jeruk No.27, RT.1/RW.9, Kb. Jeruk, Kec. Kb. Jeruk, Jakarta Barat, Indonesia, 11530
*e-mail: ekamiranda@binus.ac.id

(***received*:** 6 Maret 2023, ***revised*:** 26 April 2023, ***accepted*:** 4 Mei 2023)

Abstrak

Penelitian ini bertujuan untuk mengimplementasi text mining dalam analisis sentimen opini masyarakat Indonesia terhadap vaksinasi COVID-19 pada media sosial Twitter dengan teknik text classification Support Vector Machine (SVM) dan Random Forest. Tahap penelitian diawali dengan crawling data dari Twitter dengan periode September 2021 sampai Oktober 2021; pembersihan data; penerjemahan teks ke Bahasa Inggris; data preprocessing dengan NTLK yang dilakukan dengan dan tanpa proses lemmatization; analisis sentimen menggunakan TextBlob; pembagian data training dan testing dengan metode Hold-Out 70:30 dan 80:20; hyperparameter tuning dengan GridSearchCV; text classification dengan SVM dan Random Forest; dan pengujian hasil klasifikasi dengan menghitung Accuracy, Precision, Recall, F-Measure berbasis confusion matrix. Hasil penelitian menunjukkan bahwa text classification Random Forest secara konsisten memiliki tingkat akurasi lebih tinggi dari SVM dengan nilai akurasi tertinggi 90,59% dan sebagian besar sentimen menunjukkan netral terhadap program vaksinasi COVID-19.

Kata kunci: analisis sentimen, *text mining*, *twitter*, *support vector machine*, *random forest*

Abstract

The purpose of this study is to implement text mining for sentiment analysis of Indonesian public opinion on COVID-19 vaccination on Twitter social media using text classification techniques Support Vector Machine (SVM) and Random Forest. The research begins with crawling data from Twitter from September 2021 to October 2021; data cleansing; text translation into English; data preprocessing using NTLK performed with and without the lemmatization process; sentiment analysis using TextBlob; distribution of training and testing data with the Hold-Out method of 70:30 and 80:20; hyperparameter tuning with GridSearchCV; text classification with SVM and Random Forest; and testing the classification results by calculating Accuracy, Precision, Recall, F-Measure based on confusion matrix. The results show that text classification Random Forest consistently has a higher accuracy rate than SVM with the highest accuracy value of 90,59% and most of the sentiments indicate neutral to the COVID-19 vaccination program.

Keywords: *sentiment analysis, text mining, twitter, support vector machine, random forest*

1 Pendahuluan

COVID-19 pertama kali ditemukan di Kota Wuhan, Provinsi Hubei, Tiongkok pada akhir Desember 2019 lalu, disebabkan oleh infeksi Virus Corona atau *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2). Virus ini sangat cepat menyerang dan menularkan siapa saja, hanya dalam rentang waktu beberapa bulan virus ini sudah menyebar ke hampir semua negara dan wilayah di dunia, termasuk Indonesia [1]. Vaksin dianggap suatu cara paling efektif dalam mengurangi proliferasi COVID-19. Pelaksanaan vaksinasi di Indonesia sudah berjalan selama sembilan bulan dan saat ini sudah memasuki tahap ketiga dan keempat. Dalam pelaksanaannya, sebagian masyarakat mendukung program vaksinasi COVID-19 ini, namun tidak sedikit yang bahkan menolak untuk diberi

 ISSN:2302-8149
e-ISSN:2540-9719

vaksin. Opini yang diutarakan masyarakat menggambarkan respon yang beragam terhadap program vaksinasi COVID-19 baik yang pro, kontra ataupun netral [2]. Berangkat dari fenomena tersebut, maka perlu untuk mengetahui preferensi sentimen masyarakat terhadap vaksinasi. Dengan begitu, pemerintah dapat mengevaluasi kebijakan mengenai pelaksanaan program vaksinasi COVID-19, baik itu dari segi informasi seperti waktu dan tempat pengadaan vaksinasi, hasil aspek vaksin dan standar usia yang diperlukan untuk vaksinasi. Dengan demikian, dapat membantu meminimalisir keresahan masyarakat terhadap pelaksanaan vaksinasi, serta menangkal berita hoax yang beredar di berbagai media, sehingga pemerintah dapat dengan mudah membangun kepercayaan dan opini publik yang positif terhadap vaksinasi.

Pemanfaatan data yang bersumber dari media sosial menjadi suatu alternatif mencari sumber data untuk menggantikan survei tradisional yang dinilai sulit dilakukan karena memerlukan fase yang relatif panjang dan waktu yang lama. Pengumpulan data melalui media sosial dinilai lebih efisien biaya dalam mengakuisisi data minimum. Data diperoleh secara *real time* dan lebih detail untuk menggambarkan opini masyarakat yang sebenarnya [3]. Demikian juga data yang bersumber dari media sosial dapat dimanfaatkan untuk mengetahui kecenderungan opini masyarakat terhadap program vaksinasi COVID-19 yang dilaksanakan pemerintah. Twitter seringkali digunakan sebagai medium untuk mengekspresikan diri maupun emosi mengenai sesuatu hal, baik mengekspresikan emosi dalam bentuk pujian maupun celaan. Dengan jumlah karakter *tweet* yang hanya mencapai 140 karakter saja, Twitter sudah menjadi tempat yang efektif bagi pengguna untuk mengekspresikan opini tentang produk, layanan atau hal lain [4]. Emosi *tweet* dari para pengguna Twitter dapat diidentifikasi melalui analisa opini atau sentimen. Analisis sentimen dilakukan untuk melihat pendapat atau kecenderungan opini terhadap sebuah objek permasalahan bagi seseorang, apakah opini tersebut memiliki sentimen positif, negatif, maupun netral.

Sentiment analysis (analisis sentimen) atau sering disebut juga dengan *opinion mining* (penambangan opini) adalah studi komputasional yang menganalisis opini/pendapat, sentimen, penilaian, sikap dan emosi orang terhadap suatu entitas (seperti produk, layanan, organisasi, individu, masalah, topik, dan atributnya) yang dituangkan dalam bentuk teks [5], [6]. Analisis sentimen dengan menganalisis konten (*content analysis*) dari *Twitter*, lebih mencerminkan kondisi yang sesungguhnya yang terjadi di lapangan, karena orang lebih bebas berpendapat di media sosial.

Penelitian berkaitan dengan analisis sentimen telah banyak dilakukan. Pada tahun 2020 terdapat penelitian berjudul "Implementasi *Text Mining* pada Analisis Sentimen Opini Masyarakat terhadap Hubungan Perdagangan Indonesia dan China dengan Teknik *Text Classification Naive Bayes* dan SVM" (Hartanto, et al., 2021). Pengujian dilakukan menggunakan VADER dan TextBlob sebagai penamaan kelas label, serta menggunakan SpaCy dan NLTK sebagai *tools data preprocessing*. Hasil akurasi tertinggi bernilai 76,4% didapat dari perbandingan data *training* dan *testing* 80:20 menggunakan teknik SVM dengan VADER dan proses *data preprocessing* menggunakan *tool* SpaCy tanpa proses *lemmatization*.

Pada tahun 2020, terdapat penelitian berkaitan dengan analisis sentimen dengan judul "Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma *Naive Bayes*, *Random Forest* dan *Support Vector Machine*" yang dilakukan oleh Fitri, Yuliani, Rosyida, dan Windu Gata. Tujuan dari penelitian tersebut adalah untuk mengetahui perasaan dan opini pengguna terhadap aplikasi belajar *online* Ruangguru. Penulis menggunakan metode *Naive Bayes*, *Random Forest*, dan *Support Vector Machine* dalam melakukan klasifikasi setiap *review* yang didapatkannya. Dari hasil pengujian dari ketiga algoritma tersebut, metode algoritma *Random Forest* mendapat nilai akurasi tertinggi dengan nilai 97,16% dan nilai AUC 0,9666, disusul algoritma *Support Vector Machine* dengan nilai akurasi 96,01% dan nilai AUC 0,543, serta algoritma *Naive Bayes* dengan akurasi 94,16% dan nilai AUC 0,999 [7].

Berdasarkan penelitian-penelitian terdahulu tentang analisis sentimen berbasis *Text Mining*, maka tujuan penelitian ini adalah untuk melakukan analisis opini masyarakat terhadap pelaksanaan vaksinasi COVID-19 di Indonesia dengan menggunakan teknik berbasis *Text Mining*. Penelitian ini akan memanfaatkan data dari media sosial Twitter berupa *tweet* berbahasa Indonesia. Penarikan data *tweet* akan menggunakan bahasa pemrograman *Python* dengan menghubungkan *Application Programming Interface* (API) yang disediakan oleh *Twitter*.

 ISSN:2302-8149
e-ISSN:2540-9719

2 Tinjauan Literatur (or Literature Review)

Metode *text classification* yang digunakan dalam penelitian ini adalah *Support Vector Machine* (SVM) dan *Random Forest*. Peneliti memilih menggunakan metode klasifikasi *Support Vector Machine* karena algoritma ini dinilai yang lebih akurat dibandingkan algoritma *Naive Bayes* [8]. Salah satu kelebihan *Support Vector Machine* yaitu dapat diimplementasikan secara mudah karena penentuan *support vector* dapat dirumuskan dalam QP *problem*, yaitu proses pemecahan masalah optimasi matematika tertentu yang melibatkan fungsi kuadrat [9]. Berdasarkan tingkat akurasi dan performanya tersebut, *Support Vector Machine* banyak digunakan untuk klasifikasi data, khususnya data teks.

Metode *Random Forest* dipilih karena dalam implementasinya, algoritma ini dapat meminimalisir *error* yang mungkin terjadi, dan dapat mengatasi data *training* dalam jumlah yang besar secara efisien sehingga dapat menghasilkan klasifikasi yang baik [10]. *Random Forest* juga dikenal metode yang efektif untuk mengestimasi *missing data*.

Library NLTK dipilih sebagai *preprocessing tools* dalam penelitian ini, dikarenakan NLTK memiliki performa terbaik dalam proses *tokenization* [11]. Penelitian tersebut berkaitan dengan pemilihan *library* NLP dalam menganalisis dokumentasi *software*, menguji beberapa *library* NLP seperti *NLTK (Natural Language Toolkit)*, Stanford CoreNLP, Google's SyntaxNet, dan SpaCy. Hasil penelitian menunjukkan NLTK memiliki nilai tertinggi dibanding *library* lainnya. NLTK memiliki nilai kesamaan sebesar 98% dengan proses tokenisasi manual.

Penelitian ini juga membandingkan hasil akurasi penggunaan algoritma SVM dan *Random Forest* dalam melakukan *text classification*. Data yang digunakan dalam penelitian ini diperoleh selama masa pandemi COVID-19 dari bulan September 2021 hingga Oktober 2021 terhadap kata kunci 'vaksinasi'. *State-of the art* yang digunakan pad apenelitian ini dapat dilihat pada Tabel 1.

Tabel 1. State of the art

Penelitian	Isi Penelitian	Referensi
Analisis sentimen dengan teknik *text classification Naive Bayes*	**Tujuan dan Metode** Menganalisis sentimen masyarakat terhadap hubungan perdagangan Indonesia dan China menggunakan bahasa pemrograman Python dengan teknik *text classification Naive Bayes* dan SVM pada media sosial Twitter dan membandingkan tingkat akurasi kedua metode tersebut. *Tools* yang digunakan dalam penelitian ini adalah VADER dan TextBlob dalam analisis sentimen, NLTK dan SpaCy dalam *data preprocessing*. **Hasil Penelitian** Hasil sentimen masyarakat cenderung menunjukan respon positif terhadap hubungan perdagangan Indonesia dan China dengan nilai akurasi tertinggi 76,40% didapat ketika menggunakan VADER untuk analisis sentimen, menggunakan *library* SpaCy tanpa *lemmatization* untuk proses *text preprocessing*, menggunakan teknik *text classification* SVM dengan perbandingan antara data *training* dan *testing* 80:20. **Saran Penulis** Penulis menyarankan menggunakan API Tweepy dalam *crawling data*. Dikarenakan API Tweepy lebih *reliable* dibanding dengan API Twint. **Alasan Menjadi Tinjauan** Berdasarkan penelitian tersebut, teknik SVM menghasilkan nilai akurasi yang cukup baik dan akan digunakan sebagai referensi. Dalam penelitian tersebut dilakukan beberapa eksplorasi dalam	[12]

 ISSN:2302-8149
e-ISSN:2540-9719

	pengolahan data seperti dilakukan dengan dan tanpa *lemmatization* pada proses *data preprocessing*. Oleh karena itu, penelitian tersebut dapat dijadikan referensi bagi penulis dalam menentukan proses pengolahan data dan membandingkannya dengan teknik klasifikasi lainnya. *Range* pada penentuan label dalam penelitian tersebut dijadikan standar *range* bagi penulis. *Range* pembagian kelas yang digunakan adalah *range* >0 untuk kelas sentimen positif, <0 untuk kelas sentimen negatif dan 0 untuk kelas sentimen netral. Pengembangan dari penelitian tersebut dengan menggunakan algoritma SVM dan *Random Forest*. Serta, menggunakan *library* Tweepy pada *crawling data*.	
Analisis sentimen menggunakan algoritma *Naive Bayes, Random Forest* dan *Support Vector Machine* (SVM)	**Tujuan dan Metode** Menganalisis sentimen terhadap *review* aplikasi belajar *online* Ruangguru dari *website* Google Play Store dengan menerapkan metode klasifikasi *Naive Bayes*, *Random Forest* dan *Support Vector Machine*. **Hasil Penelitian** Hasil analisis sentimen menunjukkan tingkat penilaian respon positif lebih tinggi daripada respon negatif pada aplikasi Ruangguru. Selain itu, dari hasil pengujian tiga algoritma didapat nilai akurasi tertinggi ada pada pengujian dengan menggunakan model algoritma *Random Forest* sebesar 97,16% serta nilai AUC (Area Under Curve) 0,996. **Alasan Menjadi Tinjauan** Berdasarkan hasil dari penelitian algoritma *Random Forest* memiliki akurasi tertinggi dibanding kedua algoritma lainnya disusul dengan algoritma SVM. Oleh karena itu, *Random Forest* dipilih sebagai algoritma pembanding dalam penelitian ini dan penelitian tersebut akan dijadikan referensi dalam membandingkan metode klasifikasi *Random Forest* dan *Support Vector Machine* dalam menganalisis sentimen terhadap vaksinasi. Selain itu, penelitian ini akan melanjutkan penelitian tersebut dengan melakukan proses *testing* dengan metode diluar *Cross Validation* seperti metode *Hold Out.*	[7]
Penjelasan hasil empiris *split data training* dan *testing* untuk klasifikasi teks	**Tujuan dan Metode** Penelitian ini bertujuan untuk menganalisis hubungan antara dataset *training* dan *testing* dalam menghasilkan keakuratan model yang optimal dengan menjelaskan hasil empiris penelitian yang melatih model menggunakan data *training* dan menggunakan data *testing* untuk mengukur tingkat akurasi model. **Hasil Penelitian** Berdasarkan studi empiris menunjukkan bahwa pembagian *dataset* terbaik diperoleh menggunakan 20-30% data untuk pengujian, dan sisanya 70-80% dari data untuk pelatihan. Dari penelitian didapat bahwa hasil pembagian terbaik didapat dengan	[13]

 ISSN:2302-8149
e-ISSN:2540-9719

	perbandingan data *training* dan *testing* 80:20 karena nilai terkecil p dengan nilai (1 − p) terbesar dicapai ketika p = 0.8. **Alasan Menjadi Tinjauan** Berdasarkan penelitian tersebut pembagian *dataset training* dan *testing* sebesar 70:30 dan 80:30 memiliki performa yang baik dalam memaksimalkan tingkat akurasi model dalam memprediksi label, sehingga hal tersebut menjadi referensi dalam pembagian *dataset training* dan *testing* untuk proses *hold out* dalam penelitian ini.	
Analisis sentimen perbandingan metode klasifikasi *Random Forest* dan SVM	**Tujuan dan Metode** Menganalisis sentimen publik mengenai penerapan PSBB dengan medium *tweet* pada media sosial Twitter. Metode yang digunakan dalam analisis sentimen menggunakan algoritma SVM dan *Random Forest.* **Hasil Penelitian** Algoritma *Support Vector Machine* dianggap lebih baik dibandingkan *Random Forest* karena mampu mengenali *tweet* dengan label "Positif". **Saran Penulis** Penulis menyarankan jumlah data yang cukup banyak dapat membantu model mengenali sentimen sebuah *tweet* lebih baik lagi. Kemudian penulis juga menyarankan *text vectorization* khusus Bahasa Indonesia akan sangat membantu pra proses data dalam memperkaya informasi *input* yang akan dilatih. **Kekurangan dalam Penelitian** Dalam penelitian ini hanya menerapkan 4 pra proses data. Pra proses data yang dilakukan hanya *cleansing, case folding, removing stopword,* dan *vectorization.* Proses lainnya tidak diterapkan seperti *tokenization, stemming,* dan *POS tagging.* **Alasan Menjadi Tinjauan** Penelitian tersebut akan digunakan sebagai referensi dalam pembelajaran dalam penggunaan dan teori mengenai algoritma *Random Forest* dan *Support Vector Machine*. Pengembangan dari penelitian tersebut adalah menambahkan proses *data preprocessing* seperti *tokenization, stemming, POS tagging* dan *lemmatization*.	[14]
Library NLP untuk klasifikasi	**Tujuan dan Metode** Penelitian ini bertujuan untuk membandingkan akurasi keempat *library Natural Language Processing* (NLP) yaitu SyntaxNet dari Google, CoreNLP dari Stanford, *library* bahasa pemrograman Python NLTK, dan SpaCy. Metode pembanding menggunakan analisis dokumentasi perangkat lunak dari *Stack Overflow, file ReadMe* GitHub, dan dokumentasi API bahasa pemrograman Java. **Hasil Penelitian** Hasil dari keempat *library* tersebut beragam. Ditemukan bahwa *library* SpaCy memiliki akurasi	[11]

 ISSN:2302-8149
e-ISSN:2540-9719

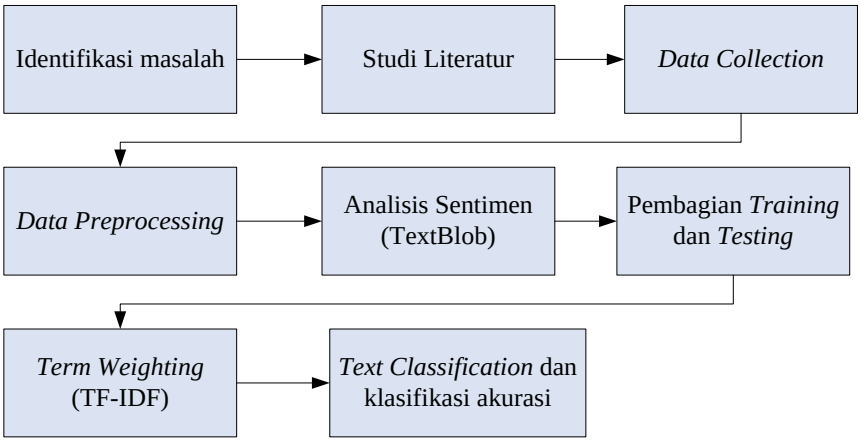
tertinggi secara keseluruhan, tetapi juga ditemukan bahwa NLTK memiliki performa terbaik dalam proses tokenisasi.

Alasan Menjadi Tinjauan Pustaka

Berdasarkan penelitian tersebut NLTK memiliki performa yang baik dalam proses *data preprocessing* terutama dalam proses tokenisasi, sehingga hal tersebut menjadi referensi dalam memilih library NLTK untuk proses *data preprocessing* dalam penelitian ini.

3 Metode Penelitian

Dalam melakukan suatu penelitian diperlukan perencanaan penelitian agar penelitian yang dilakukan dapat berjalan dengan baik, sistematis serta efektif. Terdapat delapan tahapan yang dilakukan pada penellitian ini. Tahapan-tahapan tersebut terdiri dari identifikasi masalah, studi literatur, pengumpulan data, persiapan data, analisis data, pembagian data training dan testing, term weighting (TF-IDF) dan text classification. Kerangka pikir penelitian ditampilkan pada Gambar 1.



Gambar 1. Kerangka Pikir

Identifikasi Masalah

Pada tahap ini peneliti mendefinisikan arah penelitian dan menentukan metode - metode yang akan digunakan dalam proses analisis sentimen. Masalah yang akan dibahas dalam penelitian ini adalah bagaimana cara mengimplementasi *Text Mining* dan mengukur tingkat akurasi hasil dari implementasi *Text Mining* untuk analisis sentimen opini masyarakat Indonesia terhadap vaksinasi COVID-19 dengan teknik *text classification Support Vector Machine* dan *Random Forest*.

Studi Literatur

Pada tahap ini peneliti mengumpulkan data dari buku, jurnal, website, dan referensi lainnya untuk dijadikan sebagai teori dasar yang mendukung dalam proses analisis data.

Pengumpulan Data

Pengumpulan data opini masyarakat Indonesia terhadap pelaksanaan vaksinasi COVID-19 dilakukan dengan menggunakan teknik *Data Crawling*. *Crawling* adalah teknik pengumpulan data yang digunakan untuk mengindeks informasi pada halaman menggunakan URL dengan menyertakan API. Untuk melakukan *Data Crawling* di Twitter diperlukan *library* Tweepy pada Python. Kata kunci yang digunakan adalah "vaksinasi". Data ditarik dari Bulan September 2021 hingga Oktober 2021, dan akan disimpan ke dalam dokumen Excel (.xlsx).

 ISSN:2302-8149
e-ISSN:2540-9719

Persiapan Data

Dari *tweet* yang terambil kemudian akan dilakukan pembersihan data (*Data Cleansing*) yang meliputi penghapusan pungtuasi (tanda baca dan simbol), penghapusan *tweet* yang duplikat dan menerjemahkan *tweet* ke Bahasa Inggris. Langkah selanjutnya akan dilakukan *preprocessing data* yang meliputi tahapan *case folding, tokenization, POS tagging,* pembersihan *tweet* terhadap unsur-unsur yang tidak dibutuhkan dalam analisis (*stopword*), melakukan *stemming* dan proses terakhir dilakukan dalam dua skenario yaitu melalui *lemmatization* dan tanpa *lemmatization*. *Library* yang digunakan dalam *Natural Language Processing* (NLP) adalah NLTK.

Analisis Data

Analisis data sentimen akan diklasifikasikan ke dalam sentimen positif, negatif dan netral menggunakan *TextBlob*. *TextBlob* merupakan salah satu *tools* atau *library* yang biasa digunakan untuk pemrosesan di bidang *Natural Language Processing* (NLP) menggunakan bahasa pemrograman Python. TextBlob dapat digunakan dalam berbagai proses terhadap kata teks mulai dari yang sederhana seperti tokenisasi (pemotongan kata), klasifikasi, frasa kata, penerjemah, analisis sentimen, dan lain sebagainya. *Tools* ini juga memberikan API yang sederhana sehingga mudah dalam mengakses aktivitas NLP [15].

Selanjutnya, analisis sentimen akan menggunakan metode pendekatan *supervised learning* dengan metode klasifikasi *Support Vector Machine* dan *Random Forest* dari *library Scikit-Learn. Confusion Matrix* akan digunakan sebagai metode perhitungan akurasi pada penelitian ini. *Confusion Matrix* akan berkaitan dengan perhitungan akurasi, *precision, recall* dan *f-measure.*

Pembagian Data *Training* dan *Testing*

Model akan diuji dengan membagi data menggunakan metode *Hold Out* yaitu memecah data menjadi data *training* dan *testing. Training data* adalah data yang digunakan untuk membuat model algoritma. Sedangkan *testing data* digunakan untuk menguji data yang menghasilkan akurasi algoritma selama proses *testing* [16]. Penelitian ini akan memecah data menjadi komposisi 80% dan 70% untuk data *training*, sedangkan 20% dan 30% untuk data *testing*. *Library* yang digunakan adalah *scikit-learn.model_selection* kelas *train_test_split.* Pada tahapan ini, peneliti juga menguji model analisis sentimen yang telah dibangun dengan dua algoritma klasifikasi yaitu *Support Vector Machine* dan *Random Forest.*

***Term Weighting* (TF-IDF)**

Sebelum proses klasifikasi teks dan kalkulasi akurasi, diperlukan transformasi *dataset* yang berbentuk teks (bertipe kategori) menjadi sebuah matriks yang berisi numerik (*encoding*). Kemudian akan dilakukan pembobotan kata menggunakan TF-IDF (*term frequency-inverse document frequency)*. TF-IDF adalah cara yang paling umum dalam menghitung bobot setiap kata yang digunakan dalam *information retrieval.*

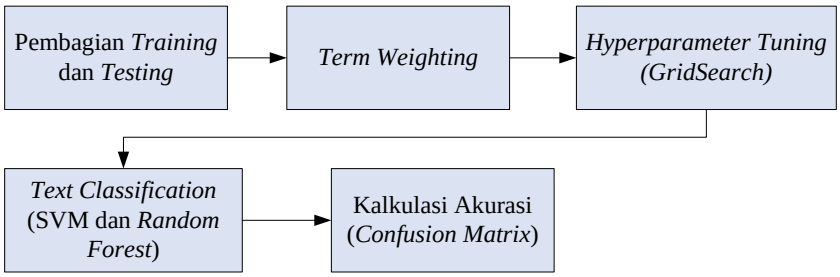
TF-IDF terdiri dari tiga faktor yaitu *term frequency, inverse document frequency* (IDF) dan *document length*. Dalam prosesnya, frekuensi akan berperan penting dalam menunjukkan seberapa umum kata tersebut dalam dokumen. Semakin banyak kata yang muncul pada dokumen, maka menunjukkan seberapa umum kata tersebut. Begitu juga dengan bobot kata, semakin sering muncul dalam suatu dokumen maka semakin besar pula bobot yang dihasilkan. Sedangkan semakin jarang kata yang muncul maka semakin kecil pula bobot yang dihasilkan.

Pada penelitian ini, pembobotan TF-IDF menggunakan kelas *TFIDFVectorizer* dari *scikit-learn*. *TFIDFVectorizer* membantu dalam menciptakan matriks TF-IDF yang menampung seluruh kata dan skor yang ada pada seluruh dokumen. Serta, menghitung kata - kata umum pada dokumen.

Text Classification

Text Classification merupakan proses pelabelan teks ke dalam kategori tertentu dari kumpulan data yang sudah ditentukan sebelumnya. Teks yang telah dikelompokkan kemudian dianalisis dengan model berbeda yang bertugas memberi tag atau kategori menurut isinya [17]. Pada penelitian ini, klasifikasi teks akan menggunakan dua *classifier* yaitu *Support Vector Machine* dan *Random Forest*. Adapun kerangka proses *Text Classification* digambarkan pada Gambar 2.

 ISSN:2302-8149
e-ISSN:2540-9719



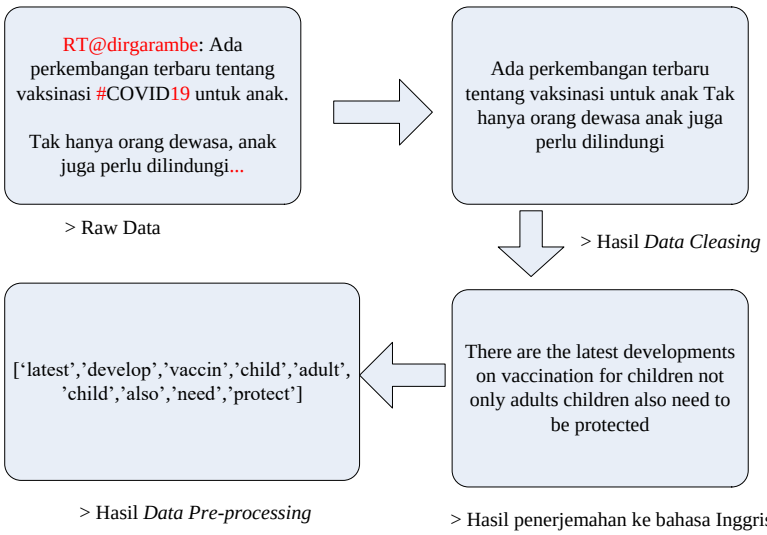
Gambar 2. Proses *Text Classification*

4 Hasil dan Pembahasan

Penelitian ini menggunakan *text classification Support Vector Machine* (SVM) dan *Random Forest*, *TextBlob* sebagai *tools* untuk melakukan analisis sentimen dan penentuan kelas label serta *library* NLTK untuk melakukan *data pre-processing*.

Hasil data collection dan pre-proses data

Hasil dari *data crawling* terkumpul sebanyak 2.500 *tweet*. Setelah *tweet* duplikat dihilangkan, total *tweet* bersih tanpa duplikat adalah sebanyak 1.415 *tweet*. *Tweet* tersebut akan melalui proses *data cleansing* sampai *data pre-processing*. Berikut ini adalah contoh hasil dari proses *data crawling* sampai *data pre-processing*. Contoh hasil proses *data crawling* hingga data *pre-processing* ditampilkan pada Gambar 3.



Gambar 3. Contoh Hasil Proses *Data Crawling* Hingga *Data Pre-Processing*

Hasil pengujian analisis *sentiment*

Hasil dari penentuan kelas label oleh *TextBlob* dengan data yang melalui proses *lemmatization* menunjukkan bahwa kelas label "netral" memiliki 946 data, kelas label "positif" berjumlah 343 data, sedangkan kelas label "negatif" berjumlah 125 data, dari keseluruhan data sebanyak 1.415 data. Sebagian hasil *TextBlob* data dengan *lematization* ditampilkan pada Gambar 4.

 ISSN:2302-8149
e-ISSN:2540-9719

	username	tweet	clean_tweet	en_tweet	text	text_tokenized	text_tagging	text_stopword	text_stemmed	text_lemmatized	polarity	label
0	pqrzan	RT @KKMPutrajaya: Selepas rentas negeri dibena...	Selepas rentas negeri dibenarkan tiada sekata...	After cross-state allowed no roadblocks to ch...	after cross-state allowed no roadblocks to ch...	['after', 'cross-state', 'allowed', 'no', 'roa...	[('after', 'IN'), ('cross-state', 'NN'), ('all...	['cross-state', 'allowed', 'roadblocks', 'chec...	['cross-stat', 'allow', 'roadblock', 'check', ...	['cross-stat', 'allow', 'roadblock', 'check', ...	0.0	netral
1	mcreativepky	Jaga Vaksinasi Covid-19 di Puskesmas Tangkilin...	Jaga Vaksinasi Covid di Puskesmas Tangkiling K...	Keep Covid Vaccination at Tangkeliling Puskesk...	keep covid vaccination at tangkeliling puskesk...	['keep', 'covid', 'vaccination', 'at', 'tangke...	[('keep', 'VB'), ('covid', 'NN'), ('vaccinatio...	['keep', 'covid', 'vaccination', 'tangkeliling...	['keep', 'covid', 'vaccin', 'tangkelil', 'pusk...	['keep', 'covid', 'vaccin', 'tangkelil', 'pusk...	0.0	netral
2	veemarisa	RT @KKMPutrajaya: mengawal pergerakan mereka y...	mengawal pergerakan mereka yang enggan meneri...	Control the movement of those who refuse to r...	control the movement of those who refuse to r...	['control', 'the', 'movement', 'of', 'those', ...	[('control', 'VB'), ('the', 'DT'), ('movement'...	['control', 'movement', 'refuse', 'receive', '...	['control', 'movement', 'refus', 'receiv', 'va...	['control', 'movement', 'refus', 'receiv', 'va...	0.0	netral
3	GSaraw4	Vaksinasi untuk Menyelamatkan diri sendiri dan...	Vaksinasi untuk Menyelamatkan diri sendiri dan...	Vaccination to save yourself and the vaccine f...	vaccination to save yourself and the vaccine f...	['vaccination', 'to', 'save', 'yourself', 'and...	[('vaccination', 'NN'), ('to', 'TO'), ('save',...	['vaccination', 'save', 'vaccine', 'family', '...	['vaccin', 'save', 'vaccin', 'famili', 'prove'...	['vaccin', 'save', 'vaccin', 'famili', 'prove'...	0.5	positif
4	Papalov03973158	Vaksinasi penting dilakukan untuk meningkatkan...	Vaksinasi penting dilakukan untuk meningkatkan...	Important vaccination is done to increase the ...	important vaccination is done to increase the ...	['important', 'vaccination', 'is', 'done', 'to...	[('important', 'JJ'), ('vaccination', 'NN'), (...	['important', 'vaccination', 'done', 'increase...	['import', 'vaccin', 'done', 'increas', 'bodi'...	['import', 'vaccin', 'done', 'increas', 'bodi'...	0.0	netral

Gambar 4. Sebagian Hasil TextBlob Data dengan Lemmatization

Hasil dari penentuan kelas label oleh *TextBlob* dengan data tanpa proses *lemmatization* menunjukkan bahwa kelas label "netral" memiliki 943 data, untuk kelas label "positif" berjumlah 341 data, sedangkan kelas label "negatif" berjumlah 130 data. Dari keseluruhan data sebanyak 1.414 data. Sebagian hasil *TextBlob* data tanpa *lematization* ditampilkan pada Gambar 5.

	username	tweet	clean_tweet	en_tweet	text	text_tokenized	text_tagging	text_stopword	text_stemmed	polarity	label
0	pqrzan	RT @KKMPutrajaya: Selepas rentas negeri dibena...	Selepas rentas negeri dibenarkan tiada sekata...	After cross-state allowed no roadblocks to ch...	after cross-state allowed no roadblocks to ch...	['after', 'cross-state', 'allowed', 'no', 'roa...	[('after', 'IN'), ('cross-state', 'NN'), ('all...	['cross-state', 'allowed', 'roadblocks', 'chec...	['cross-stat', 'allow', 'roadblock', 'check', ...	0.0	netral
1	mcreativepky	Jaga Vaksinasi Covid-19 di Puskesmas Tangkilin...	Jaga Vaksinasi Covid di Puskesmas Tangkiling K...	Keep Covid Vaccination at Tangkeliling Puskesk...	keep covid vaccination at tangkeliling puskesk...	['keep', 'covid', 'vaccination', 'at', 'tangke...	[('keep', 'VB'), ('covid', 'NN'), ('vaccinatio...	['keep', 'covid', 'vaccination', 'tangkelilling...	['keep', 'covid', 'vaccin', 'tangkelil', 'pusk...	0.0	netral
2	veemarisa	RT @KKMPutrajaya: mengawal pergerakan mereka y...	mengawal pergerakan mereka yang enggan meneri...	Control the movement of those who refuse to r...	control the movement of those who refuse to r...	['control', 'the', 'movement', 'of', 'those', ...	[('control', 'VB'), ('the', 'DT'), ('movement'...	['control', 'movement', 'refuse', 'receive', '...	['control', 'movement', 'refus', 'receiv', 'va...	0.0	netral
3	GSaraw4	Vaksinasi untuk Menyelamatkan diri sendiri dan...	Vaksinasi untuk Menyelamatkan diri sendiri dan...	Vaccination to save yourself and the vaccine f...	vaccination to save yourself and the vaccine f...	['vaccination', 'to', 'save', 'yourself', 'and...	[('vaccination', 'NN'), ('to', 'TO'), ('save',...	['vaccination', 'save', 'vaccine', 'family', '...	['vaccin', 'save', 'vaccin', 'famili', 'prove'...	0.5	positif
4	Papalov03973158	Vaksinasi penting dilakukan untuk meningkatkan...	Vaksinasi penting dilakukan untuk meningkatkan...	Important vaccination is done to increase the ...	important vaccination is done to increase the ...	['important', 'vaccination', 'is', 'done', 'to...	[('important', 'JJ'), ('vaccination', 'NN'), (...	['important', 'vaccination', 'done', 'increase...	['import', 'vaccin', 'done', 'increas', 'bodi'...	0.0	netral

Gambar 05. Sebagian Hasil TextBlob Data Tanpa Lemmatization

Hasil akurasi *text classification*

Tabel 2 menempilkan hasil akurasi SVM dan *Random Forest.*

Tabel 2. Hasil akurasi SVM dan Random Forest (RF)

Lematization	***Training-Testing 70:30***		***Training-Testing 80:20***	
	SVM	***RF***	***SVM***	***RF***
Ya	84.24%	90,59%	84,45%	88,34%
Tidak	84%	87,29%	85,51%	90,11%

Dari perbandingan data *training* dan *testing* 70:30 ditemukan bahwa nilai akurasi tertinggi untuk algoritma *Support Vector Machine* adalah 84,24% dengan data yang diperlakukan melalui proses *lemmatization.* Sedangkan nilai akurasi tertinggi *Random Forest* adalah 90,59% dari data yang diperlakukan melalui proses *lemmatization*. Nilai akurasi terendah *Support Vector Machine* adalah 84% dengan data diperlakukan tanpa *lemmatization*. Untuk nilai akurasi terendah *Random Forest* adalah 87,29% dengan perlakuan data tanpa *lemmatization*.

Sementara itu untuk perbandingan data *training* dan *testing* 80:20 ditemukan bahwa nilai akurasi tertinggi diperoleh *Support Vector Machine* dengan nilai 85,51% dengan data yang diperlakukan

 ISSN:2302-8149
e-ISSN:2540-9719

tanpa melalui proses *lemmatization.* Sedangkan nilai akurasi tertinggi *Random Forest* dengan nilai 90,11% dari data yang diperlakukan tanpa *lemmatization*. Nilai akurasi terendah *Support Vector Machine* adalah 84,45% dengan data melalui proses *lemmatization*. Untuk nilai akurasi terendah *Random Forest* adalah 88,34% dengan perlakuan data melalui proses *lemmatization.*

Dapat dilihat bahwa data yang melalui proses *lemmatization* memiliki nilai akurasi yang lebih tinggi pada pembagian data *training* 70% dan data *testing* 30%. Sedangkan data tanpa melalui proses *lemmatization* memiliki nilai akurasi lebih tinggi pada pembagian data *training* 80% dan data *testing* 20%. Selain itu, dari hasil penelitian juga dapat dilihat secara keseluruhan bahwa algoritma *Random Forest* memiliki nilai akurasi lebih tinggi dibandingkan dengan *Support Vector Machine* (SVM).

Tabel 3 menempilkan hasil *precision* SVM dan *Random Forest.*

Tabel 3. Hasil precision SVM dan Random Forest (RF)

Lematization	***Training-Testing 70:30***		***Training-Testing 80:20***	
	SVM	***RF***	***SVM***	***RF***
Ya	83,6%	90,77%	84,67%	89,02%
Tidak	83,69%	87,76%	84,94%	90,53%

Dari perbandingan data *training* dan *testing* 70:30 ditemukan bahwa nilai *precision* tertinggi untuk algoritma *Support Vector Machine* adalah 83,69% dengan data yang diperlakukan tanpa melalui proses *lemmatization.* Sedangkan nilai *precision* tertinggi *Random Forest* adalah 90,77% dari data yang diperlakukan melalui proses *lemmatization*. Nilai *precision* terendah *Support Vector Machine* adalah 83,6% dengan data diperlakukan melalui proses *lemmatization*. Untuk *Random Forest* nilai terendah dengan nilai 87,76% dengan perlakuan data tanpa *lemmatization*.

Sementara itu untuk perbandingan data *training* dan *testing* 80:20 ditemukan bahwa nilai *precision* tertinggi diperoleh *Support Vector Machine* dengan nilai 84,94% dengan data yang diperlakukan tanpa melalui proses *lemmatization.* Sedangkan nilai *precision* tertinggi *Random Forest* dengan nilai 90,53% dari data yang diperlakukan tanpa *lemmatization*. Nilai *precision* terendah *Support Vector Machine* adalah 84,67% dengan data diperlakukan *lemmatization*. Untuk nilai *precision* terendah *Random Forest* adalah 89,02% dengan perlakuan data melalui proses *lemmatization.*

Dapat dilihat secara keseluruhan bahwa algoritma *Random Forest* memiliki nilai *precision* lebih tinggi dibandingkan dengan *Support Vector Machine* (SVM). Algoritma *Random Forest* mencapai nilai *precision* tertinggi pada pembagian data *training* 70% dan data *testing* 30% dengan data *lemmatization*, sedangkan pada pembagian data *training* 80% dan data *testing* 20% nilai *precision* lebih tinggi dengan data tanpa *lemmatization*. Tabel 4 menempilkan hasil *recall* SVM dan *Random Forest.*

Tabel 4. Hasil recall SVM dan Random Forest (RF)

Lematization	***Training-Testing 70:30***		***Training-Testing 80:20***	
	SVM	***RF***	***SVM***	***RF***
Ya	84,24%	90,59%	84,45%	88,34%
Tidak	84%	87,29%	85,51%	90,11%

Dari perbandingan data *training* dan *testing* 70:30 ditemukan bahwa nilai *recall* tertinggi untuk algoritma *Support Vector Machine* adalah 84,24% dengan data yang diperlakukan melalui proses *lemmatization.* Sedangkan nilai *recall* tertinggi *Random Forest* adalah 90,59% dari data yang diperlakukan dengan *lemmatization*. Nilai *recall* terendah *Support Vector Machine* adalah 84% dengan data diperlakukan tanpa melalui proses *lemmatization*. Untuk nilai *recall* terendah *Random Forest* adalah 87,29% dengan perlakuan data tanpa *lemmatization.*

Sementara itu untuk perbandingan data *training* dan *testing* 80:20 ditemukan bahwa nilai *recall* tertinggi diperoleh *Support Vector Machine* dengan nilai 85,51% dengan data yang diperlakukan tanpa melalui proses *lemmatization.* Sedangkan nilai *recall* tertinggi *Random Forest* adalah 90,11%

 ISSN:2302-8149
e-ISSN:2540-9719

dari data yang diperlakukan tanpa *lemmatization*. Nilai *recall* terendah *Support Vector Machine* adalah 84,45% dengan data melalui proses *lemmatization*. Untuk *Random Forest* nilai *recall* terendah dengan nilai 88,34% dengan perlakuan data melalui proses *lemmatization*.

Dapat dilihat bahwa data yang melalui proses *lemmatization* memiliki nilai *recall* yang lebih tinggi pada pembagian data *training* 70% dan data *testing* 30%. Sedangkan data tanpa melalui proses *lemmatization* memiliki nilai *recall* lebih tinggi pada pembagian data *training* 80% dan data *testing* 20%. Selain itu, dari hasil penelitian juga dapat dilihat secara keseluruhan bahwa algoritma *Random Forest* memiliki nilai *recall* lebih tinggi dibandingkan dengan *Support Vector Machine* (SVM).

Tabel 5 menempilkan hasil *f-score* SVM dan *Random Forest.*

Tabel 5. Hasil f-score SVM dan Random Forest (RF)

Lematization	***Training-Testing 70:30***		***Training-Testing 80:20***	
	SVM	***RF***	***SVM***	***RF***
Ya	82,1%	89,99%	82,8%	87,41%
Tidak	82,34%	85,96%	84,33%	89,31%

Dari perbandingan data *training* dan *testing* 70:30 ditemukan bahwa nilai *f-score* tertinggi untuk algoritma *Support Vector Machine* adalah 82,34% dengan data yang diperlakukan tanpa melalui proses *lemmatization.* Sedangkan nilai *f-score* tertinggi *Random Forest* adalah 89,99% dari data yang diperlakukan melalui proses *lemmatization*. Nilai *f-score* terendah *Support Vector Machine* adalah 82,21% dengan data diperlakukan melalui proses *lemmatization*. Untuk nilai *f-score* terendah *Random Forest* adalah 85,96% dengan perlakuan data tanpa *lemmatization.*

Sementara itu untuk perbandingan data *training* dan *testing* 80:20 ditemukan bahwa nilai *f-score* tertinggi diperoleh *Support Vector Machine* dengan nilai 84,33% dengan data yang diperlakukan tanpa melalui proses *lemmatization.* Sedangkan nilai *f-score* tertinggi *Random Forest* dengan nilai 89,31% dari data yang diperlakukan tanpa *lemmatization*. Nilai *f-score* terendah *Support Vector Machine* adalah 82,8% dengan data melalui proses *lemmatization*. Untuk nilai *f-score* terendah *Random Forest* adalah 87,41% dengan perlakuan data melalui proses *lemmatization.*

Dapat dilihat secara keseluruhan bahwa algoritma *Random Forest* memiliki nilai *f-score* lebih tinggi dibandingkan dengan *Support Vector Machine* (SVM). Algoritma *Random Forest* mencapai nilai *f-score* tertinggi pada pembagian data *training* 70% dan data *testing* 30% dengan data *lemmatization*, sedangkan pada pembagian data *training* 80% dan data *testing* 20% nilai *f-score* lebih tinggi dengan data tanpa *lemmatization*.

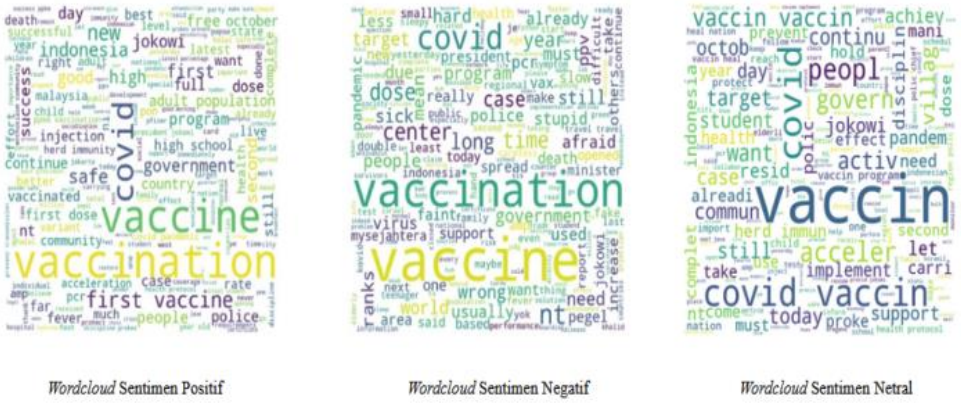
Ada beberapa temuan yang diperoleh penelitian sebagai bahan evaluasi. Dalam proses data *crawling tweet* peneliti menggunakan *library* Tweepy. Tweepy hanya dapat melakukan penarikan data maksimal 10 hari terakhir. Oleh karena itu penulis perlu melakukan berulang kali penarikan dalam selang waktu 10 hari tiap penarikan. Selain itu, terdapat beberapa *tweet* tidak secara lengkap terambil pada proses tersebut dikarenakan adanya fitur *"read more"* yang membuat *tweet* yang ditarik hanya berupa "...".

Kemudian hal yang perlu menjadi bahan evaluasi selanjutnya adalah pada tahap *preprocessing*. Dimana data yang melalui proses *lemmatization* tidak menunjukkan perbedaan signifikan, dikarenakan proses *lemmatization* menggunakan kata yang sudah dilalui proses *stemming*. Hal ini berdampak pada pendefinisian kata dasar yang dilakukan oleh *lemmatization*. Sehingga *lemmatization* hanya mengulang definisi kata dasar yang telah didefinisikan oleh *stemming.*

Pada penelitian yang dilakukan hasil analisis sentimen dengan bantuan anotasi *TextBlob* ditemukan bahwa data yang diperlakukan dengan *lemmatization* memiliki nilai positif sebesar 343 *tweet*, netral sebesar 946 tweet dan negatif sebesar 125 *tweet*. Sedangkan proses tanpa *lemmatization* nilai positif sebesar 341 *tweet*, netral sebesar 943 *tweet* dan negatif sebesar 130 *tweet*. Kemudian hasil ini juga digunakan kembali untuk proses selanjutnya pada klasifikasi teks pada kedua algoritma yaitu *Support Vector Machine* dan *Random Forest*.

 ISSN:2302-8149
e-ISSN:2540-9719

Berdasarkan pengelompokan sentimen didapat hasil visualisasi *wordcloud* yang menggambarkan kumpulan kata terbanyak dalam analisis teks untuk sentimen positif dan negatif sebagai berikut (lihat Gambar 6).



Gambar 6. Wordcloud untuk Sentimen Positif, Negatif dan Netral

Berdasarkan hasil *wordcloud*, representasi kata untuk *tweet* bersentimen positif digambarkan oleh kata '*first vaccine', 'safe', 'success',* dan *'government'*. Sedangkan untuk *tweet* sentimen negatif kebanyakan direpresentasikan oleh kata '*long', 'time', 'dose', 'sick', 'center',* dan *'afraid'*. Kemudian untuk sentiment netral digambarkan oleh kata '*acceleration*', '*continue*', dan '*reached*'. Untuk kata '*vaccination', 'vacccine',* dan *'covid'* karena menggambarkan keseluruhan *tweet* baik itu untuk bersentimen positif, negatif, ataupun netral, maka dapat diabaikan. Selain itu, pengambilan *tweet* menggunakan *keyword* "vaksinasi", membuat kata "*vaccination*" dan "*vaccine*" menjadi kata dominan diseluruh visualisasi sentimen.

Kemudian hasil dari pengelompokan sentimen juga digunakan kembali untuk proses selanjutnya pada klasifikasi teks menggunakan kedua algoritma yaitu *Support Vector Machine* dan *Random Forest*. Pada kedua algoritma tersebut menunjukkan hasil prediksi sebagai berikut (lihat Tabel 6).

Tabel 6. Hasil f-score SVM dan Random Forest (RF)

Training:Testing	*Lematization*	SVM			RF		
		(+)	(-)	(netral)	(+)	(-)	(netral)
70:30	Ya	67	284	7	79	287	19
	Tidak	67	279	11	71	286	14
80:20	Ya	42	188	9	50	188	12
	Tidak	43	192	7	47	198	10

Pada perbandingan data *training* dan *testing* 70:30 ditemukan bahwa data yang diperlakukan melalui proses *lemmatization* dengan algoritma *Support Vector Machine* berhasil memprediksi sentimen positif sebesar 67 data, sentimen netral dengan 284 data dan sentimen negatif sebesar 7 data. Sedangkan proses tanpa *lemmatization* ditemukan sentimen positif sebesar 67 data, sentimen netral sebesar 279 data dan sentimen negatif sebesar 11 data.

Perbandingan data *training* dan *testing* 80:20 ditemukan bahwa data yang diperlakukan melalui proses *lemmatization* dengan algoritma *Support Vector Machine* berhasil memprediksi sentimen positif sebesar 42 data, sentimen netral dengan 188 data dan sentimen negatif sebesar 9 data. Sedangkan proses tanpa *lemmatization* ditemukan sentimen positif sebesar 43 data, sentimen netral sebesar 192 data dan sentimen negatif sebesar 7 data.

Pada perbandingan data *training* dan *testing* 70:30 ditemukan bahwa data yang diperlakukan melalui proses *lemmatization* dengan algoritma *Random Forest* berhasil memprediksi sentimen positif

 ISSN:2302-8149
e-ISSN:2540-9719

sebesar 79 data, sentimen netral dengan 287 data dan sentimen negatif sebesar 19 data. Sedangkan proses tanpa *lemmatization* ditemukan sentimen positif sebesar 71 data, sentimen netral sebesar 286 data dan sentimen negatif sebesar 14 data.

Perbandingan data *training* dan *testing* 80:20 ditemukan bahwa data yang diperlakukan melalui proses *lemmatization* dengan algoritma *Random Forest* berhasil memprediksi sentimen positif sebesar 50 data, sentimen netral dengan 188 data dan sentimen negatif sebesar 12 data. Sedangkan proses tanpa *lemmatization* ditemukan sentimen positif sebesar 47 data, sentimen netral sebesar 198 data dan sentimen negatif sebesar 10 data.

Berdasarkan hasil analisis sentimen dan prediksi dari kedua algoritma tersebut menunjukkan sentimen netral memiliki jumlah yang paling banyak dibandingkan kedua nilai sentimen lainnya. Hal ini menyatakan bahwa masyarakat Indonesia pada periode tersebut yaitu bulan September hingga Oktober 2021 memiliki pendapat yang cenderung netral mengenai program vaksinasi yang diselenggarakan pemerintah. Hal ini dikarenakan sebagian besar masyarakat Indonesia telah menerima vaksinasi COVID-19.

5 Kesimpulan

Penelitian yang dilakukan bertujuan untuk mengimplementasikan *text mining* dalam analisis sentimen opini masyarakat Indonesia terhadap pelaksanaan vaksinasi COVID-19 di Indonesia dengan teknik *text classification Support Vector Machine* dan *Random Forest* serta mengukur tingkat akurasi kedua algoritma tersebut. Pada penelitian ini, objek penelitian menggunakan data dari Twitter berupa *tweet* berbahasa Indonesia. *Tweet* yang diambil merupakan *tweet* yang berkaitan dengan pelaksanaan vaksinasi COVID-19, dan kata kunci yang digunakan dalam proses *crawling* adalah "vaksinasi". Penarikan data dilakukan dalam kurun waktu bulan September 2021 hingga Oktober 2021 yang berjumlah sebanyak 2.500 *tweet*.

Hasil penelitian menunjukkan bahwa pada kedua algoritma yang dipakai yaitu *Support Vector Machine* dan *Random Forest*, terdapat perbedaan tingkat akurasi yang cukup signifikan. *Random Forest* memiliki nilai akurasi lebih tinggi sebesar 4,53% dari *Support Vector Machine*. Dari keseluruhan percobaan baik percobaan yang dilakukan dengan pembagian data *training-testing* 70:30 maupun 80:20, dan proses pengolahan data dengan *lemmatization* maupun tanpa *lemmatization*, ditemukan bahwa *Random Forest* secara konsisten memiliki tingkat akurasi lebih tinggi dari *Support Vector Machine*, dengan tingkat akurasi tertinggi sebesar 90,59% yang didapat dengan data melalui proses *lemmatization* dan pembagian data *training-testing* sebesar 70:30. Oleh karena itu, kedua hipotesis nol (H0) telah terpenuhi, yaitu terdapat perbedaan tingkat akurasi antara kedua algoritma tersebut dan tingkat akurasi *Random Forest* lebih tinggi dari *Support Vector Machine*. Dari hasil penelitian ini juga dapat disimpulkan pengaruh pemilihan nilai *parameter* dalam proses *hyperparameter tuning* mempengaruhi akurasi kedua algoritma tersebut.

Berdasarkan penelitian analisis sentimen ini menunjukkan hasil bahwa sentimen netral memperoleh nilai tertinggi dibanding sentimen lainnya yaitu dengan nilai persentase 67%. Disusul persentase sentimen positif sebesar 24% dan persentase sentimen negatif sebesar 9%. Hal ini menunjukkan bahwa sebagian besar masyarakat Indonesia pada periode tersebut memiliki pendapat yang cenderung netral mengenai pelaksanaan vaksinasi COVID-19 yang diselenggarakan oleh pemerintah.

Adapun saran dari yang dapat digunakan untuk penelitian yang akan datang sebagai berikut: menggunakan *library* NLP lainnya dalam proses *data preprocessing* seperti VADER, *CoreNLP*, *SyntaxNet* sebagai pembanding; melakukan pengubahan kata Bahasa Indonesia tidak baku menjadi kata baku, dengan menggunakan kata baku dapat memberikan analisis sentimen yang berbeda; dalam proses *lemmatization* menggunakan kata yang telah dilakukan penghapusan *stopword*, dibandingkan menggunakan kata dari hasil proses *stemming*; dalam proses pemberian label dapat mengganti jumlah pemberian label sentimen yang digunakan, misalnya menjadi dua (positif dan negatif) ataupun lima label sentiment, hal ini bertujuan untuk mengetahui apakah dengan mengganti jumlah pemberian label tersebut dapat menimbulkan sudut pandang yang berbeda dalam menentukan sentimen yang terkandung; menggunakan metode *K-Fold* dalam pembagian data *training* dan *testing* sebagai pembanding; menggunakan algoritma lainnya seperti *Naive Bayes classifier*, *K-Nearest Neighbor* ataupun *Decision Tree* sebagai pembanding.

 ISSN:2302-8149
e-ISSN:2540-9719

Referensi

[1] S. Sumengen, S. Sagala, B. Sutomo, W. Liem, H. Al Hamid, "Strengthening the Strategic and Operational Response for Reducing COVID-19 Transmission in Indonesia," Jurnal Kesehatan Masyarakat Nasional., vol. 16, no. 1, pp. 3-10, 2021.
[2] Litbangkes, "Tantangan Pelaksanaan Vaksinasi COVID-19 di Indonesia," Badan Penelitian dan Pengembangan Kesehatan, 2021. [Online]. Available: https://www.litbang.kemkes.go.id/tantangan-pelaksanaan-vaksinasi-covid-19-di-indonesia/. [Accessed: April. 15, 2022]
[3] F. Rachman, "Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter," Indonesian of Health Information Management Journal., vol. 8, no. 2, pp. 2655–9129, 2020.
[4] K. Perdana, T. Pricillia, "Optimasi TextBlob Menggunakan Support Vector Machine untuk Analisis Sentimen (Studi Kasus Layanan Telkomsel)," Bangkit Indonesia., vol. 10, no. 1, pp. 13–15, 2021.
[5] L. Zhang, B. Liu, "Sentiment Analysis and Opinion Mining," Encyclopedia of Machine Learning and Data Mining., pp. 1152–1161, 2017.
[6] A. Prasanti, M. A. Fauzi, M. T. Furqon, "Klasifikasi Teks Pengaduan Pada Sambat Online Menggunakan Metode N- Gram dan Neighbor Weighted K-Nearest Neighbor (NW-KNN)," Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, vol. 2, no. 2, pp. 594–601, 2018.
[7] E. Fitri, Y. Yuliani, S. Rosyida, W. Gata, "Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes, Random Forest Dan Support Vector Machine,",Jurnal Transformatika., vol. 18, no. 1, pp.. 71-80, 2020.
[8] K. V. S. Toy, Y. A. Sari, I. Cholissodin, "Analisis Sentimen Twitter menggunakan Metode Naive Bayes dengan Relevance Frequency Feature Selection (Studi Kasus: Opini Masyarakat mengenai Kebijakan New Normal)," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer., vol. 5, no. 11, pp. 5068-5074, 2021.
[9] V. I. Santoso, G. Virginia, Y, Lukito, "Penerapan Sentiment Analysis Pada Hasil Evaluasi Dosen Dengan Metode Support Vector Machine," Jurnal Transformatika, vol. 14, no. 2, pp. 79-82, 2017
[10] S. Z. Wenno, "Pendekatan Random Forest Pada Pohon Klasifikasi Dan Multivariate Adaptive Regression Spline Untuk Keakuratan Klasifikasi Pengguna Narkoba Di Jawa Timur", Thesis Sarjana, Universitas Airlangga, Surabaya, 2017.
[11] F. N. A. Al Omran, C. Treude, "Choosing an NLP Library for Analyzing Software Documentation: A Systematic Literature Review and a Series of Experiments," Proceedings IEEE/ACM 14th International Conference on Mining Software Repositories, pp. 187–197, 2017.
[12] N. Hartanto, A. A. B. Raharjo, A. G. Pambudhi, "Implementasi Text Mining Pada Analisis Sentimen Opini Masyarakat Terhadap Hubungan Perdagangan Indonesia Dan China Dengan Teknik Text Classification Naive Bayes", Tesis Sarjana, Universitas Bina Nusantara, Jakarta, 2021.
[13] B. Genç, H. Tunç, "Optimal training and Test Sets Design for Machine Learning," Turkish Journal of Electrical Engineering & Computer Sciences, vol. 27, pp. 1534 – 1545, 2019.
[14] M. R. Adrian, M. P. Putra, M. H. Rafialdy, N. A. Rakhmawati, N. A, "Perbandingan Metode Klasifikasi Random Forest dan SVM Pada Analisis Sentimen PSBB," Jurnal Informatika Upgris., vol. 7, no. 1, pp. 36–40, 2021.
[15] S. Dewi, D. B. Arianto, "Twitter Sentiment Analysis Towards Qatar As Host of The 2022 World Cup using Textblob, " Journal of Social Research, vol. 2, no. 2, pp. 443-454, 2023.
[16] A. Fatoni, "Optimasi Aplikasi Antrian Pasien Online Menggunakan Algoritma Patient Treatment Time Prediction", Thesis Sarjana, Universitas 17 Agustus, Jakarta, 2020.
[17] K. Shah, H. Patel, D. Sanghvi, M. Shah, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification,"Augmented Human Research, vol. 5, no.1, pp. 1- 12, 2020.