**IMPLEMENTASI TEXT MINING PADA ANALISIS SENTIMEN DENGAN TEXT CLASSIFICATION SUPPORT VECTOR MACHINE DAN RANDOM FOREST**

**Veronica Gabriella**Universitas Bina Nusantara

School of Information SystemsJl. Kebon Jeruk Raya No.27, Kebon Jeruk, Jakarta Barat, 11530, (021) 5345830  
veronica.gabriella@binus.ac.id **Sriyanda Afrida Wahyudi**Universitas Bina Nusantara

School of Information SystemsJl. Kebon Jeruk Raya No.27, Kebon Jeruk, Jakarta Barat, 11530, (021) 5345830sriyanda.wahyudi@binus.ac.id **Jennifer Chai**Universitas Bina Nusantara

School of Information SystemsJl. Kebon Jeruk Raya No.27, Kebon Jeruk, Jakarta Barat, 11530, (021) 5345830jennifer.chai@binus.ac.id **Eka Miranda**Universitas Bina Nusantara

School of Information SystemsJl. Kebon Jeruk Raya No.27, Kebon Jeruk, Jakarta Barat, 11530, (021) 5345830ekamiranda@binus.ac.id

**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.*

***Keyword:*** *sentiment analysis, text mining, Twitter, SVM, Random Forest*

## *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, SVM, *Random Forest*.

**PENDAHULUAN**

Saat ini dunia sedang dihadapi sebuah tantangan besar yaitu pandemi COVID-19 (*Coronavirus Disease* 2019). Penyakit 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 (Susilo et al., 2020).

Berdasarkan data dari situs *Worldometer* per tanggal 14 September 2021, tercatat dalam 233 negara, kasus positif COVID-19 yang dilaporkan mencapai total sebanyak 226,217,805 kasus. Di wilayah Indonesia sendiri, angka kasus positif COVID-19 tercatat sebanyak 4,174,216 kasus dengan jumlah kematian mencapai 139,415 orang. Melihat angka kasus yang semakin meningkat, beragam upaya pun dilakukan berbagai negara termasuk Indonesia dengan memberlakukan berbagai kebijakan khusus untuk menekan penyebaran COVID-19.

Upaya yang dilakukan oleh pemerintah Indonesia meliputi penyiagaan rumah sakit rujukan, mengadakan *Rapid Test Antigen,* PCR (*Polymerase Chain Reaction)* dan melakukan pemasangan alat pendeteksi suhu tubuh di tempat umum (seperti bandar udara, stasiun kereta api, dll). Kemudian pemerintah juga menerapkan Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM), memperketat disiplin protokol kesehatan demi menekan angka penularan COVID-19, hingga pelaksanaan vaksinasi demi tercapainya *herd immunity*/kekebalan kelompok.

Vaksin dianggap suatu cara paling efektif dalam mengurangi proliferasi COVID-19. Oleh karena itu, pemerintah berupaya untuk mendatangkan vaksin dari beberapa negara untuk mengatasi masalah COVID-19, sesuai yang tercantum dalam Peraturan Presiden (Perpres) Nomor 99 Tahun 2020 membahas Pengadaan Vaksin dan Pelaksanaan Vaksinasi dalam rangka Penanggulangan Pandemi COVID-19, yang ditandatangani pada 5 Oktober 2020. Dalam pelaksanaannya, Indonesia sudah mendatangkan beberapa vaksin seperti vaksin Sinovac (RRT), Sinopharm (RRT), AstraZeneca (Inggris), Moderna (Amerika Serikat), dan Pfizer (Jerman-Amerika Serikat).

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 vaksin. Berdasarkan hasil survei Katadata Insight Center (2021) menyatakan sebesar 6,1% masyarakat enggan melengkapi vaksinasi kedua walaupun sudah disuntik dosis pertama. Sebesar 56% meragukan efektifitas dan keampuhan vaksin COVID-19, sementara 23,7% responden menilai efek samping vaksin terlalu berat. Selama periode tersebut, terdapat berbagai macam opini masyarakat Indonesia terhadap jalannya pelaksanaan vaksinasi yang diungkapkan melalui berbagai media. Salah satu media yang banyak digunakan oleh masyarakat untuk menyalurkan pendapatnya adalah media sosial.

Opini yang diutarakan masyarakat menggambarkan respon yang beragam terhadap program vaksinasi COVID-19 baik yang pro, kontra ataupun netral (Litbangkes, 2021). Adanya penolakan dari masyarakat terhadap pelaksanaan vaksinasi, dapat menghalangi tujuan pemerintah Indonesia dalam melaksanakan vaksinasi terhadap seluruh masyarakat Indonesia.

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 (Rachman et al., 2020). Demikian juga data yang bersumber dari media sosial dapat dimanfaatkan untuk mengetahui kecenderungan opini masyarakat terhadap program vaksinasi COVID-19 yang dilaksanakan pemerintah.

Dari data yang dihimpun dalam *Digital Report 2021*, jumlah pengguna media sosial di Indonesia hingga saat ini mencapai lebih dari 170 juta pengguna. Salah satu media sosial yang banyak digunakan oleh masyarakat Indonesia adalah Twitter, yang mencakup 63,6% total pengguna media sosial di Indonesia (Simon Kemp, 2021).

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 (Perdana & Pricillia, 2021). 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.

Analisis sentimen termasuk dalam teknik penambangan data berupa teks atau dikenal dengan istilah *Text Mining*. *Text Mining* adalah penambangan teks menggunakan seperangkat alat analisis untuk menghasilkan suatu informasi baru yang belum diketahui sebelumnya atau menemukan kembali informasi yang tersirat secara implisit. Informasi didapat dari hasil seleksi otomatis sumber-sumber data teks yang berbeda-beda melalui identifikasi dan eksplorasi pola yang menarik (Prasanti et al., 2018).

*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 (Zhang & Liu, 2017). 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 (Fitri et al., 2020).

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.

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* (Sudiantoro & Zuliarso, 2018)*.* 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 (Nur & Fithriasari, 2016)*.* Berdasarkan tingkat akurasi dan performanya tersebut, *Support Vector Machine* banyak digunakan untuk klasifikasi data, khususnya data teks.

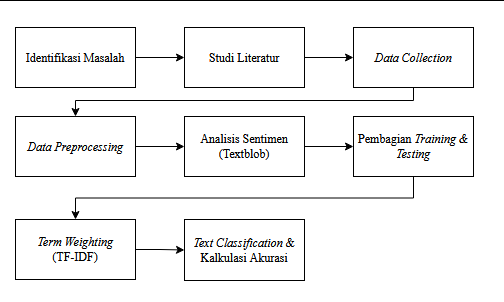
Metode *Random Forest* dipilih karena dalam implementasinya, algoritma ini dapat meminimalisir *error* yang mungkin terjadi, dan dapat mengatasi data *training* dalam jumlah yang besar secara efisien sehingga dapat menghasilkan klasifikasi yang baik (Wenno, 2017). *Random Forest* juga dikenal metode yang efektif untuk mengestimasi *missing data*.

*Library* NLTK dipilih sebagai *preprocessing tools* dalam penelitian ini, dikarenakan NLTK memiliki performa terbaik dalam proses *tokenization* (Al Omran & Treude, 2017). 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.

Berdasarkan latar belakang tersebut, penulis tertarik untuk melakukan penelitian dengan judul “Implementasi *Text Mining* pada Analisis Sentimen Opini Masyarakat Indonesia terhadap Vaksinasi COVID-19 pada Media Sosial Twitter dengan *Text Classification Support Vector Machine* dan *Random Forest*”. Penelitian ini dilakukan untuk mengetahui bagaimana respon dan opini masyarakat Indonesia terhadap pelaksanaan vaksinasi COVID-19 yang sudah berjalan sejak awal tahun 2021 dengan menggunakan data yang bersumber dari media sosial Twitter. Penelitian menggunakan *tools* analisis sentimen TextBlob serta *library* NLP yaitu NLTK. 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’.

**METODE PENELITIAN**

    Berikut ini adalah tahapan metode penelitian ini:



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).

#### 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

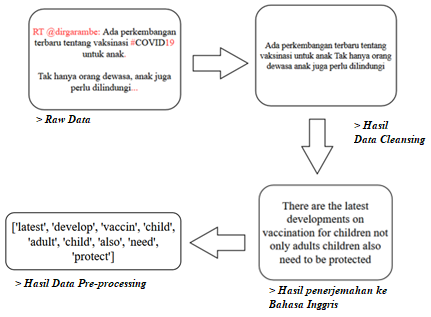
#### Analisis data sentimen akan diklasifikasikan ke dalam sentimen positif, negatif dan netral menggunakan TextBlob. Selanjutnya, analisis sentimen akan menggunakan metode pendekatan supervised learning dengan metode klasifikasi Support Vector Machine dan Random Forest dari library Scikit-Learn. Confusion Matrix akan digunakan sebagai metode perhitungan akurasi pada penelitian ini. Confusion Matrix akan berkaitan dengan perhitungan akurasi, precision, recall dan f-measure.

**HASIL PENELITIAN**

    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*.

1. **Hasil *data collection* dan pre-proses data**

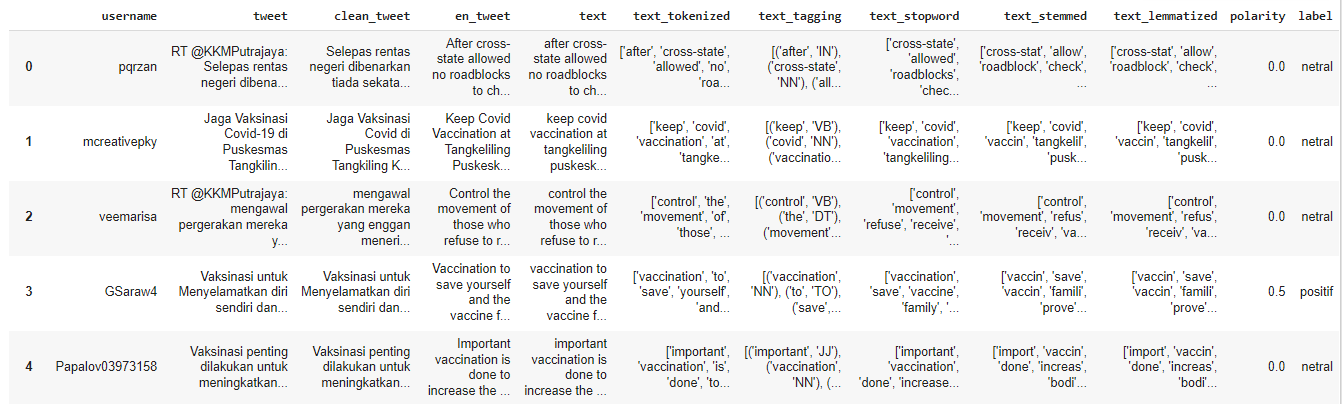
Hasil dari *data crawling* terkumpul sebanyak 2.500 *tweet*. Setelah *tweet* duplikat dihilangkan, total *tweet* bersih tanpa duplikat adalah sebanyak 1.415 *tweet*. *Tweet* tersebut akan melalui proses *data* *cleansing* sampai *data* *pre-processing*. Berikut ini adalah contoh hasil dari proses *data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



Gambar 2 Contoh Hasil Proses Data Crawling Hingga Data Pre-processing

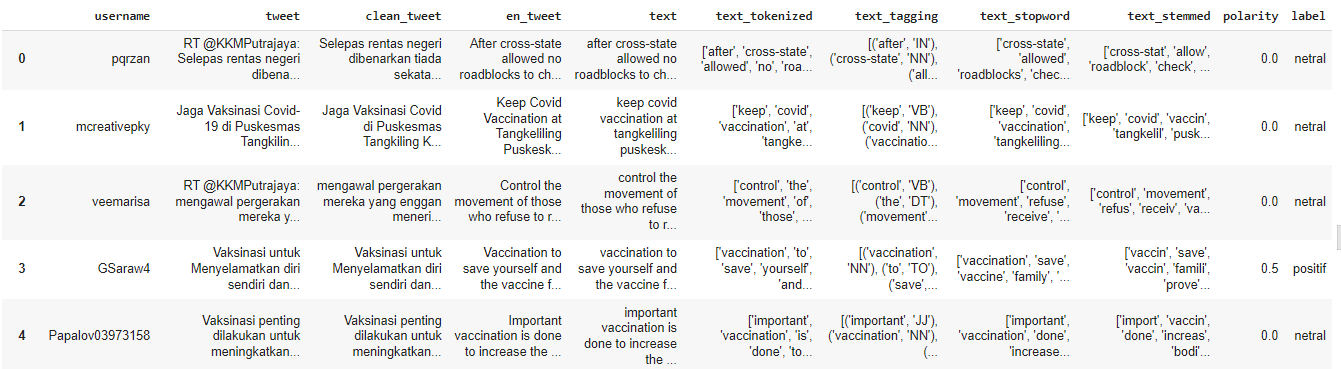
1. **Hasil pengujian analisis sentiment**

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.



Gambar 3 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.



Gambar 4 Sebagian Hasil TextBlob Data Tanpa Lemmatization

1. **Hasil akurasi *text classification***

Berikut ini adalah table hasil akurasi SVM dan *Random Forest*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Perlakuan** | | | ***Training-Testing* 70:30** | | ***Training-Testing* 80:20** | |
| **Analisis sentimen** | ***Data***  ***Pre-processing*** | ***Lemma-tization*** | **SVM** | ***Random Forest*** | **SVM** | ***Random Forest*** |
| TextBlob | NLTK | Ya | 84,24 | 90,59 | 84,45 | 88,34 |
| TextBlob | NLTK | Tidak | 84 | 87,29 | 85,51 | 90,11 |
| **Rata-rata** | | | **84,12** | **88,94** | **84,98** | **89,225** |

Tabel 1 Hasil Akurasi SVM dan Random Forest

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 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.*

**PEMBAHASAN**

1. **Analisis hasil *data collection* dan pre-proses data**

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.*

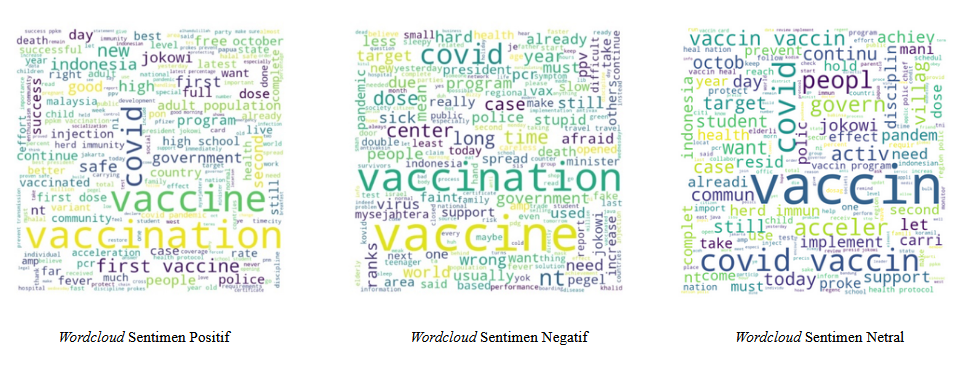
1. **Analisis hasil analisis sentimen**

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*. Pada kedua algoritma tersebut menunjukkan hasil prediksi sebagai berikut.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Perlakuan** |  | ***Training-Testing* 70:30** | | ***Training-Testing* 80:20** | |
| ***Lemma-tization*** | ***Label*** | **SVM** | ***Random Forest*** | **SVM** | ***Random Forest*** |
| Ya | Positif | 67 | 79 | 42 | 50 |
| Ya  Ya  Tidak  Tidak  Tidak | Netral  Negatif  Positif  Netral  Negatif | 284  7  67  279  11 | 287  19  71  286  14 | 188  9  43  192  7 | 188  12  47  198  10 |

Tabel 2 Hasil Prediksi SVM dan Random Forest

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



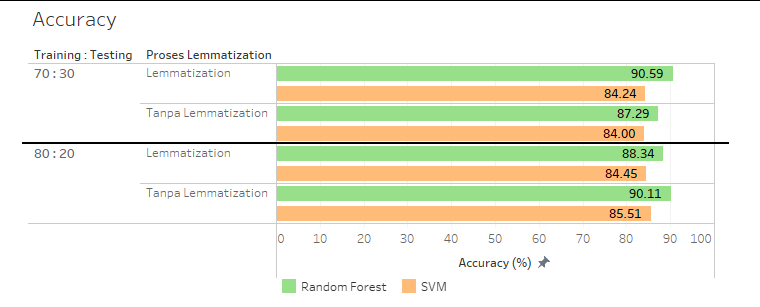
Gambar 5 Wordcloud untuk Sentimen Positif, Negatif dan Netral

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.

1. **Analisis hasil akurasi *text-classification***

Dapat dilihat bahwabaik pada pembagian data *training-testing* 70:30 maupun 80:20, hasil akurasi *text-classification* diperoleh dari data dengan penamaan kelas label menggunakan *TextBlob* dan proses *data pre-processing* dengan *library* NLTK.

Kemudian dari garfik dibawah 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%.



Gambar 6 Hasil Akurasi SVM dan Random Forest

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).

**SIMPULAN DAN SARAN**

    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 penulis yang dapat digunakan untuk penelitian yang akan datang sebagai berikut:

1. Penulis menyarankan agar menggunakan *library* NLPlainnya dalam proses *data preprocessing* seperti VADER, *CoreNLP, SyntaxNet* sebagai pembanding.
2. Penulis menyarankan agar melakukan pengubahan kata Bahasa Indonesia tidak baku menjadi kata baku. Dengan menggunakan kata baku dapat memberikan analisis sentimen yang berbeda.
3. Penulis menyarankan agar dalam proses *lemmatization* menggunakan kata yang telah dilakukan penghapusan *stopword,* dibandingkan menggunakan kata dari hasil proses *stemming*.
4. Penulis menyarankan untuk penelitian selanjutnya dalam proses pemberian label dapat mengganti jumlah pemberian label sentimen yang digunakan, misalnya menjadi dua (positif dan negatif) ataupun lima label sentimen. Hal ini bertujuan untuk mengetahui apakah dengan mengganti jumlah pemberian label tersebut dapat menimbulkan sudut pandang yang berbeda dalam menentukan sentimen yang terkandung.
5. Penulis menyarankan untuk penelitian selanjutnya dapat menggunakan metode *K-Fold* dalam pembagian data *training* dan *testing* sebagai pembanding.
6. Penulis menyarankan untuk menggunakan algoritma lainnya seperti *Naive Bayes classifier,* *K-Nearest Neighbor* ataupun *Decision Tree* sebagai pembanding.

**REFERENSI**

Al Omran, F. N. A., & Treude, C. (2017). Choosing an NLP Library for Analyzing Software Documentation: A Systematic Literature Review and a Series of Experiments. *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*, 187–197. https://doi.org/10.1109/MSR.2017.42

Annur, C. M. (2021, September 22). *Meragukan Efektivitas, Alasan Utama Masyarakat Enggan Vaksin Covid-19 Dosis Kedua | Databoks*. Katadata Insight Center. https://databoks.katadata.co.id/datapublish/2021/09/22/meragukan-efektivitas-alasan-utama-masyarakat-enggan-vaksin-covid-19-dosis-kedua

Fitri, E., Yuliani, Y., Rosyida, S., & Gata, W. (2020). Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes, Random Forest Dan Support Vector Machine. *Jurnal Transformatika*, *18*(1), 71. https://doi.org/10.26623/transformatika.v18i1.2317

Hartanto, N., Raharjo, A. A. B., & Pambudhi, A. G. (2021). *IMPLEMENTASI TEXT MINING PADA ANALISIS SENTIMEN OPINI MASYARAKAT TERHADAP HUBUNGAN PERDAGANGAN INDONESIA DAN CHINA DENGAN TEKNIK TEXT CLASSIFICATION NAIVE BAYES*. Bina Nusantara.

Litbangkes. (2021, March 27). *Tantangan Pelaksanaan Vaksinasi COVID-19 di Indonesia | Badan Penelitian dan Pengembangan Kesehatan*. https://www.litbang.kemkes.go.id/tantangan-pelaksanaan-vaksinasi-covid-19-di-indonesia/

Nur, S., & Fithriasari, K. (2016). *Klasifikasi Berita Online Menggunakan Metode Support Vector Machine dan K- Nearest Neighbor*. *5*(2).

Perdana, K., & Pricillia, T. (2021). *Optimasi TextBlob Menggunakan Support Vector Machine untuk Analisis Sentimen ( Studi Kasus Layanan Telkomsel )*. *X*(01), 13–15.

Prasanti, A. A., Fauzi, M. A., & Furqon, M. T. (2018). 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 (J-PTIIK) Universitas Brawijaya*, *Vol. 2*(2), 594–601.

Rachman, F., Health, S. P.-I. of, & 2020, undefined. (2020). Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter. *Inohim.Esaunggul.Ac.Id*, *8*(2), 2655–9129.

Simon Kemp. (2021, February 11). *Digital in Indonesia: All the Statistics You Need in 2021*. https://datareportal.com/reports/digital-2021-indonesia

Sudiantoro, A. V., & Zuliarso, E. (2018). *Analisis Sentimen Twitter Menggunakan Text Mining Dengan Algoritma Naive Bayes Classifier*. *10*(2), 398–401.

Susilo, A., Rumende, C. M., Pitoyo, C. W., Santoso, W. D., Yulianti, M., Herikurniawan, H., Sinto, R., Singh, G., Nainggolan, L., Nelwan, E. J., Chen, L. K., Widhani, A., Wijaya, E., Wicaksana, B., Maksum, M., Annisa, F., Jasirwan, C. O. M., & Yunihastuti, E. (2020). *Coronavirus Disease 2019: Tinjauan Literatur Terkini*. Jurnal Penyakit Dalam Indonesia. https://doi.org/10.7454/jpdi.v7i1.415

Wenno, S. Z. (2017). *PENDEKATAN RANDOM FOREST PADA POHON KLASIFIKASI DAN MULTIVARIATE ADAPTIVE REGRESSION SPLINE UNTUK KEAKURATAN KLASIFIKASI PENGGUNA NARKOBA DI JAWA TIMUR*. Universitas Airlangga.

Worldometer. (2021). *COVID Live Update 14 September 2021 - Worldometer*. https://www.worldometers.info/coronavirus/

Zhang, L., & Liu, B. (2017). Sentiment Analysis and Opinion Mining. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning and Data Mining* (pp. 1152–1161). Springer US. https://doi.org/10.1007/978-1-4899-7687-1\_907

**RIWAYAT PENULIS**

**Veronica Gabriella** lahir di kota Bekasi pada tanggal 8 Juli 2000. Penulis menamatkan pendidikan S1 di Universitas Bina Nusantara dalam bidang ilmu Sistem Informasi pada tahun 2022.

**Sriyanda Afrida Wahyudi** lahir di kota Jakarta pada tanggal 12 Juni 2000. Penulis menamatkan pendidikan S1 di Universitas Bina Nusantara dalam bidang ilmu Sistem Informasi pada tahun 2022.

**Jennifer Chai** lahir di kota Jambi pada tanggal 4 Oktober 2000. Penulis menamatkan pendidikan S1 di Universitas Bina Nusantara dalam bidang ilmu Sistem Informasi pada tahun 2022.