

Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naïve Bayes

by Artikel 6

Submission date: 14-Mar-2024 10:12AM (UTC+0700)

Submission ID: 2319966654

File name: OVID-19_Menggunakan_Recurrent_Neural_Network_dan_Na_ve_Bayes.pdf (392.23K)

Word count: 4659

Character count: 26397



3
Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naïve Bayes

18 Merinda Lestandy¹, Abdurrahim Abdurrahim², Lailis Syafa'ah³

¹Program Studi D3 Teknologi Elektronika, Fakultas Teknik, Universitas Muhammadiyah Malang

^{2,3}Program Studi Teknik Elektro, Fakultas Teknik, Universitas Muhammadiyah Malang

¹merindalestandy@umm.ac.id, ²abdurrahik@webmail.umm.ac.id, ³lailis@umm.ac.id

Abstract

COVID-19 has become a global pandemic including Indonesia, so the government is taking vaccinations as a preventive measure. The public's response to this continues to appear on social media platforms, one of which is Twitter. Tweets about the COVID-19 vaccine have generated various kinds of positive and negative opinions in the community. Therefore, it is very important to detect and filter it to prevent the spread of incorrect information. Sentiment analysis is a method used to determine the content of a dataset in the form of negative, positive or neutral text. The dataset in this study was obtained from 5000 COVID-19 vaccine tweets with the distribution of 3800 positive sentiment tweets, 800 negative sentiment tweets and 400 neutral sentiment tweets. The dataset obtained is then pre-processed data to optimize data processing. There are 4 stages of pre-processing, including remove punctuation, case folding, stemming and tokenizing. This study examines the performance of RNN and Naïve Bayes by adding the TF-IDF (Term Frequency-Inverse Document Frequency) technique which aims to give weight to the word relationship (term) of a document. The test results show that RNN (TF-IDF) has a greater accuracy of 97.77% compared to Naïve Bayes (TF-IDF) of 80%.

Keywords: Sentiment Analysis, Vaccine COVID-19, TF-IDF, RNN, Naïve Bayes

Abstrak

COVID-19 telah menjadi pandemik dunia termasuk Indonesia, sehingga pemerintah mengambil tindakan vaksinasi sebagai upaya pencegahan. Tanggapan masyarakat mengenai hal tersebut terus muncul di platform media sosial salah satunya twitter. Tweet mengenai vaksin COVID-19 menimbulkan berbagai macam opini positif dan negatif di masyarakat. Oleh karena itu, sangat penting untuk mendeteksi dan menyaringnya agar tidak terjadi penyebaran informasi yang tidak benar. Analisis sentimen merupakan suatu metode yang digunakan untuk menentukan isi dari suatu dataset dalam bentuk teks yang bersifat negatif, positif atau netral. Dataset pada penelitian ini diperoleh dari 5000 tweet vaksin COVID-19 dengan pembagian 3800 tweet sentimen positif, 800 tweet sentimen negatif dan 400 tweet sentimen netral. Dataset yang didapat kemudian dilakukan pre-processing data untuk mengoptimalkan pengolahan data. Terdapat 4 tahapan pre-processing antara lain remove punctuation, case folding, stemming dan tokenizing. Penelitian ini mengkaji kinerja RNN dan Naïve Bayes dengan menambahkan teknik TF-IDF (Term Frequency-Inverse Document Frequency) yang bertujuan untuk memberikan bobot pada hubungan kata (term) sebuah dokumen. Hasil pengujian menunjukkan RNN (TF-IDF) memiliki akurasi lebih besar yaitu 97.77% dibandingkan Naïve Bayes (TF-IDF) sebesar 80%.

Kata kunci: Analisis Sentimen, Vaksin COVID-19, TF-IDF, RNN, Naïve Bayes

1. Pendahuluan

COVID-19 (Coronavirus Disease-2019) merupakan penyakit yang disebabkan oleh virus corona jenis baru yaitu Sars-CoV-2. Virus ini menyebar dengan kontak ataupun melalui tetesan udara seperti batuk dan bersin. Berdasarkan data www.covid19.go.id, penyebaran virus COVID-19 di Indonesia per tanggal 1 Maret 2020 sampai dengan 13 Agustus 2021 sebanyak 3.804.943 terkonfirmasi positif COVID-19 dan 115.096 meninggal

akibat virus tersebut. Krisis akibat COVID-19 saat ini menciptakan situasi sosial bagi kesehatan mental masyarakat[1].

Mayoritas orang mengkonsumsi berita dari media sosial untuk pertama kalinya, dibandingkan sumber tradisional lainnya seperti televisi, surat kabar dll [2],[3]. Orang cenderung percaya apa yang mereka temukan di jejaring sosial, yang membuat mereka rentan terhadap rumor dan berita palsu. Oleh karena itu, sangat penting untuk

mendeteksi dan menyaringnya agar tidak terjadi penyebaran informasi yang tidak benar. Demikian pula informasi wabah virus COVID-19 yang terjadi di Indonesia dari bulan maret 2020 yang terus muncul di platform media sosial salah satunya *twitter*.

Analisis sentimen adalah suatu proses yang bertujuan untuk mengetahui apakah polaritas suatu data berupa teks (dokumen, kalimat, paragraf) akan mengarah ke positif, negatif, atau netral[4]. Klasifikasi teks dari postingan media sosial [5], [6] selalu menjadi masalah penelitian yang menarik dan memiliki tantangan tertentu. Kajian analisis sentimen media sosial tentang COVID-19 menghasilkan lima tema relevan yang berkisar dari positif hingga negatif[7]–[11].

Salah satu upaya tindakan pemerintah akibat persebaran COVID-19 yang begitu cepat yaitu memberikan vaksinasi kepada masyarakat sebagai upaya pencegahan[12], [13]. *Tweet* mengenai vaksin COVID-19 menimbulkan berbagai macam opini di masyarakat seperti mendukung program vaksin ataupun menolak untuk divaksin[14]–[16]. Analisis sentimen merupakan suatu metode yang digunakan untuk menentukan isi dari suatu dataset dalam bentuk teks yang bersifat negatif, positif atau netral[17]–[19].

Penelitian analisis sentimen menggunakan *machine learning* banyak dilakukan beberapa tahun terakhir [20][21]. Pengujian [20] melakukan survey mengenai *machine learning* dengan dataset yang digunakan pada penelitian ini adalah dataset dari website ulasan amazon *e-commerce*. Metodologi yang digunakan adalah data akuisisi, *pre-processing* data serta proses *training* menggunakan algoritma Naïve Bayes dan Support Vector Machine. Penelitian [21] menggunakan proses *pre-processing* yaitu *stopwords*, *stemming* dan *lemmatization* serta metode yang digunakan adalah Naïve Bayes 80,61%, Logistic Regression 82,47%, dan Support Vector Machine 83,71%.

Recurrent Neural Network (RNN) adalah salah satu arsitektur yang paling populer digunakan dalam *Natural Language Processing* (NLP) karena struktur *recurrent* cocok untuk pemrosesan teks[5]. Salah satu metode *deep learning* yang diusulkan dalam penelitian ini adalah RNN dengan penerapan arsitektur *Long Short-Term Memory* (LSTM). RNN dapat menggunakan representasi kata terdistribusi dengan terlebih dahulu mengubah token yang terdiri dari setiap teks menjadi vektor yang membentuk matriks. Selain itu, dalam pengaturan *supervised learning*, data latih dilengkapi dengan label tertentu. Namun, standar *vectorizer* seperti *Count vectorizer* tidak memperhitungkan informasi label saat membuat vektor teks. Teknik TF-IDF (*Term Frequency-Inverse Document Frequency*) bertujuan untuk mengatasi keterbatasan tersebut dengan memberi bobot pada hubungan sebuah kata (*term*) pada sebuah dokumen[22].

Penelitian [23] melakukan analisis sentimen untuk mengetahui sentimen negatif dan positif terhadap COVID-19 pada *twitter*. Metode *pre-processing* yang digunakan adalah *data cleansing*, tokenisasi kemudian *tweet* COVID-19 akan dilakukan proses *lemmatized* dan *stemmed* serta metode pembobotan kata yang digunakan pada penelitian tersebut adalah *bag of words* dan *doc2vec*. Penelitian tersebut menghasilkan akurasi sebesar 81% untuk metode *Deep Learning* lalu metode *Gaussian Fuzzy* menghasilkan akurasi 79%.

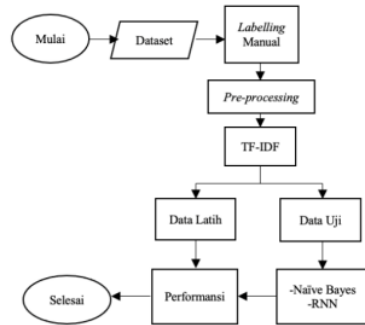
Pada Tabel 1 kami menyajikan ulasan tentang algoritma yang paling umum digunakan untuk klasifikasi sentimen. Dari tinjauan tersebut, dapat dilihat bahwa sebagian besar algoritma dapat digunakan untuk menyelesaikan masalah klasifikasi sentiment. Dalam makalah ini, kami menganalisis kinerja RNN dan Naïve Bayes dengan menambahkan teknik TF-IDF dalam menyelesaikan klasifikasi sentiment *tweet* vaksin COVID-19.

Tabel 1. Perbandingan Pendekatan Analisis Sentimen

| Referensi | Deskripsi Dataset | Metode | Akurasi |
|---|--|---|------------------------|
| A. M. Rahat, A. Kahir, and A. K. M. Masum [24] | Dataset yang digunakan yaitu <i>tweet Airline Reviews</i> yang terdiri dari ulasan positif dan negatif dengan total 10.000 data. | Support Vector Machine (SVM), Naïve Bayes | 82.48%, 76.56% |
| M. Wongkar and A. Angdresey [25] | Data dikumpulkan melalui media sosial <i>twitter</i> terkait dengan pasangan calon presiden Republik Indonesia periode 2019 – 2024. Data tersebut berjumlah 443 dengan atribut sentimen yang berisi informasi positif dan negatif. | Naïve Bayes, SVM, K-Nearest Neighbor (KNN), | 75.58%, 63.99%, 73.34% |
| G. A. Ruz, P. A. Henríquez, and A. Mascareño [26] | Dataset berisi kumpulan 2187 <i>tweet</i> dari gempa Chili 2010 yang diposting sebelum dan sesudah peristiwa kritis (27-02-2010 03:34:08). | Naïve Bayes, SVM, Random Forest | 74.2%, 81.2%, 72.5% |
| L. Kurniasari and A. Setyanto [27] | Dataset yang digunakan adalah dataset yang berisi ulasan dalam bahasa Indonesia dari situs Traveloka. Model yang akan digunakan untuk mengklasifikasikan ulasan pengguna menjadi dua kategori, ulasan positif dan negatif. | RNN | 91.9% |
| B. N. Saha, A. Senapati, and A. Mahajan [28] | Data dikumpulkan dari surat kabar Bengali "Ananda bazar Patrika" terkait dengan berita politik dalam konteks pemilu India. | RNN | 85% |
| M. F. Wahid, M. J. Hasan, and M. S. Alom [29] | Datasets "ABSA" [30] yang berisi komentar terkait kriket untuk analisis sentimen kriket dari teks Bangla. Dataset terdiri dari 2979 data dengan label positif, negatif dan netral. | RNN | 95% |

2. Metode Penelitian

Pada Gambar 1 terdapat beberapa proses yang dilakukan dalam penelitian ini, yaitu mengumpulkan dataset, proses *labelling* secara manual, *pre-processing* data, pembobotan TF-IDF, klasifikasi Naïve Bayes dan RNN, serta pengujian metode menggunakan *confusion matrix*.



Gambar 1. Metode Penelitian

2.1. Dataset

Data vaksin COVID-19 berjumlah 5000 data diambil dari <https://www.kaggle.com/rpnugroho/indonesian-vaccination-tweets/code> berdasarkan *hashtag* #vaksin dan #vaksinasi. Pengumpulan data menggunakan *tweepy* dimulai pada tanggal 10 Januari 2021 sampai dengan 21 April 2021. Dari data tersebut kemudian dilakukan pelabelan secara manual dengan 3 pembagian klasifikasi yakni 3800 *tweet* sentimen positif, 800 *tweet* sentimen negatif dan 400 *tweet* sentimen netral. Berikut adalah hasil dari label sentimen *tweet* vaksin COVID-19 ditunjukkan pada Tabel 2.

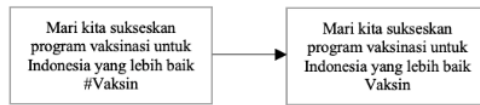
Tabel 2. Pelabelan Manual Sentimen

| Teks | Label |
|---|---------|
| Mari kita sukses kan program vaksinasi untuk Indonesia yang lebih baik #Vaksin | Positif |
| Vaksinasi Tenaga Kesehatan di Surabaya Segera Rampung #vaksin | Netral |
| Masih banyak keraguan dan penolakan di kalangan masyarakat terhadap vaksinasi Covid19 | Negatif |

2.2. Pre-Processing

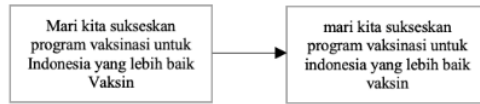
Pada tahap ini dilakukan *pre-processing* data untuk mengoptimalkan pengolahan data. Terdapat 4 tahapan *pre-processing* data pada penelitian ini antara lain *remove punctuation*, *case folding*, *stemming* dan *tokenizing*.

Tahapan pertama yaitu *remove punctuation*. *Remove Punctuation* merupakan teknik penghilangan tanda baca yang digunakan dalam sebuah teks untuk membedakan antara kalimat dan bagian penyusunnya dan untuk memperjelas maknanya[31]. Data yang telah dikumpulkan dan diberi label secara manual akan dibersihkan agar lebih mudah diproses. Pada Gambar 2 ditunjukkan contoh dari proses *remove punctuation*.



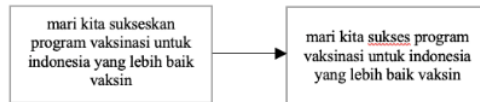
Gambar 2. Remove Punctuation

Tahapan kedua yaitu *Case Folding* seperti pada Gambar 3. *Case folding* merupakan langkah dalam pengolahan data yang bertujuan untuk mengubah atau menghilangkan semua huruf kapital pada dokumen menjadi huruf kecil [32].



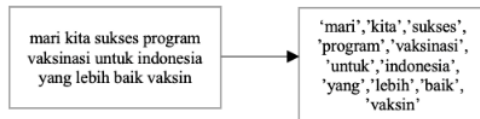
Gambar 3. Case Folding

Tahapan selanjutnya adalah *stemming*. *Stemming* adalah proses menghilangkan awalan dan akhiran dalam sebuah kata untuk mendapatkan akar kata dari sebuah dokumen[24]. Hasil dari langkah sebelumnya diolah untuk dilakukan *stemming* seperti yang ditunjukkan pada Gambar 4.



Gambar 4. Stemming

Pada Gambar 5 merupakan tahapan terakhir pada proses *pre-processing* yaitu *tokenizing*. *Tokenizing* adalah proses memecah urutan karakter menjadi beberapa bagian (kata/frasa) yang disebut token[33].



Gambar 5. Tokenizing

2.3. TF-IDF

Metode TF-IDF adalah cara untuk memberi bobot pada hubungan sebuah kata (*term*) dengan sebuah dokumen. Metode ini menggabungkan dua konsep untuk menghitung bobot, yaitu frekuensi kemunculan suatu kata dalam dokumen tertentu dan frekuensi kebalikan dari dokumen yang mengandung kata tersebut. Frekuensi kemunculan kata dalam dokumen yang disediakan menunjukkan betapa pentingnya kata tersebut dalam dokumen. Frekuensi dokumen yang berisi kata tersebut menunjukkan seberapa umum kata tersebut. Jadi, bobot hubungan antara sebuah kata dan dokumen akan tinggi jika frekuensi kata dalam dokumen juga tinggi dan frekuensi keseluruhan dokumen yang mengandung kata rendah dalam dokumen[22], [34]. Secara matematis, bobot TF-IDF dapat dihitung menggunakan Persamaan 1.

DOI: <https://doi.org/10.29207/resti.v5i4.3308>

Creative Commons Attribution 4.0 International License (CC BY 4.0)

$$TF.IDF_{std}(t) = tf_d^t \times \log \frac{N}{df_t} \quad (1)$$

Dimana tf_d^t menunjukkan berapa kali istilah t muncul dalam dokumen d . N menunjukkan jumlah total dokumen dalam korpus. df^t menunjukkan jumlah dokumen di mana istilah t terjadi. Pada Tabel 3 menunjukkan pembobotan sentimen *tweet* vaksin COVID-19.

Tabel 3. Hasil Perhitungan TF*IDF

| Term | TF | | | df | IDF | | | TF*IDF | | |
|------------|----|----|----|----|-------|-------|-------|--------|-------|-------|
| | t1 | t2 | t3 | | t1 | t2 | t3 | t1 | t2 | t3 |
| mari | 1 | 0 | 0 | 1 | 0,477 | 0,477 | 0 | 0 | 0,477 | 0 |
| kita | 1 | 0 | 0 | 1 | 0,477 | 0,477 | 0 | 0 | 0,477 | 0 |
| sukses | 1 | 0 | 0 | 1 | 0,477 | 0,477 | 0 | 0 | 0,477 | 0 |
| program | 1 | 0 | 0 | 1 | 0,477 | 0,477 | 0 | 0 | 0,477 | 0 |
| vaksinasi | 1 | 1 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| untuk | 1 | 0 | 0 | 1 | 0,477 | 0,477 | 0 | 0 | 0,477 | 0 |
| indonesia | 1 | 0 | 0 | 1 | 0,477 | 0,477 | 0 | 0 | 0,477 | 0 |
| y ang | 1 | 0 | 0 | 1 | 0,477 | 0,477 | 0 | 0 | 0,477 | 0 |
| lebih | 1 | 0 | 0 | 1 | 0,477 | 0,477 | 0 | 0 | 0,477 | 0 |
| baik | 1 | 0 | 0 | 1 | 0,477 | 0,477 | 0 | 0 | 0,477 | 0 |
| vaksin | 1 | 1 | 0 | 2 | 0,176 | 0,477 | 0 | 0,477 | 0,176 | 0,477 |
| tenaga | 0 | 1 | 0 | 1 | 0,477 | 0 | 0,477 | 0 | 0,477 | 0 |
| sehat | 0 | 1 | 0 | 1 | 0,477 | 0 | 0,477 | 0 | 0,477 | 0 |
| di | 0 | 1 | 1 | 2 | 0,176 | 0 | 0,477 | 0 | 0,176 | 0,477 |
| surabaya | 0 | 1 | 0 | 1 | 0,477 | 0 | 0,477 | 0 | 0,477 | 0 |
| segera | 0 | 1 | 0 | 1 | 0,477 | 0 | 0,477 | 0 | 0,477 | 0 |
| rampung | 0 | 1 | 0 | 1 | 0,477 | 0 | 0,477 | 0 | 0,477 | 0 |
| masih | 0 | 0 | 1 | 1 | 0,477 | 0 | 0 | 0,477 | 0 | 0,477 |
| banyak | 0 | 0 | 1 | 1 | 0,477 | 0 | 0 | 0 | 0,477 | 0 |
| ragu | 0 | 0 | 1 | 1 | 0,477 | 0 | 0 | 0 | 0,477 | 0 |
| dan | 0 | 0 | 1 | 1 | 0,477 | 0 | 0 | 0 | 0,477 | 0 |
| tolak | 0 | 0 | 1 | 1 | 0,477 | 0 | 0 | 0 | 0,477 | 0 |
| kalangan | 0 | 0 | 1 | 1 | 0,477 | 0 | 0 | 0 | 0,477 | 0 |
| masyarakat | 0 | 0 | 1 | 1 | 0,477 | 0 | 0 | 0 | 0,477 | 0 |
| hadap | 0 | 0 | 1 | 1 | 0,477 | 0 | 0 | 0 | 0,477 | 0 |
| covid19 | 0 | 0 | 1 | 1 | 0,477 | 0 | 0 | 0,477 | 0 | 0,477 |

2.4. RNN

RNN adalah jenis jaringan saraf dengan status memori untuk memproses beberapa input. Aktivasi *recurrent node* terdiri dari umpan balik untuk dirinya sendiri dari satu langkah waktu ke langkah berikutnya. RNN termasuk dalam kategori *deep learning* karena data diproses secara otomatis dan tanpa mendefinisikan fitur [26]. RNN dapat menggunakan keadaan internal (memori) untuk memproses urutan input. Hal ini berlaku dalam (NLP) [15], pengenalan suara (*speech recognition*) [25], *music synthesis* [27], dan pemrosesan data keuangan [28]. Perhitungan RNN ditunjukkan pada Persamaan 2 dan Persamaan 3 dibawah ini.

$$s_t = \tanh(U_{x_t} + W_{s_{t-1}}) \quad (2)$$

$$\hat{y}_t = \text{softmax}(V_{s_t}) \quad (3)$$

2.5. Naïve Bayes

Naïve Bayes adalah kumpulan algoritma klasifikasi yang didasarkan pada Teorema Bayes. Teknik pengklasifikasi Naïve Bayes menjadi metode yang sangat populer dalam proses sentimen analisis[24], [35], [36]. Pengklasifikasi Naïve Bayes digunakan sebagai

pengklasifikasi probabilistik. Probabilitas P didefinisikan seperti pada Persamaan 4.

$$P(H|X) = \frac{P(H)P(X)}{P(X)} \quad (4)$$

Dimana $P(H|X)$ merupakan peluang hipotesa H berdasarkan kondisi X , X adalah data latih dengan *class* (label) yang diketahui, H adalah data dengan *class* (label), $P(H)$ merupakan peluang dari hipotesa X , $P(X)$ peluang dari X yang diamati dan $P(H|X)$ adalah peluang X berdasarkan kondisi pada hipotesa H .

2.6. Performansi

Pengukuran kinerja algoritma klasifikasi pada penelitian ini yaitu dengan menggunakan *confusion matrix*. *Confusion matrix* menunjukkan hasil identifikasi antara jumlah data prediksi yang benar dan jumlah data prediksi yang salah dibandingkan dengan fakta yang dihasilkan[22]. Tabel *confusion matrix* ditunjukkan pada Tabel 4.

Tabel 4. Confusion Matrix

| Aktual | Prediksi | | |
|---------|----------|---------|---------|
| | Positif | Positif | Negatif |
| | | TP | FN |
| Negatif | FP | TN | |

Dimana:

TP (*True Positive*): data positif sentimen vaksin COVID-19 yang diprediksi dengan benar.

TN (*True Negative*): data negatif sentimen vaksin COVID-19 yang diprediksi dengan benar.

FN (*False Negative*): data positif sentimen vaksin COVID-19 yang diprediksi sebagai data negatif sentimen vaksin COVID-19.

FP (*False Positive*): data negative sentimen vaksin COVID-19 yang diprediksi sebagai data positif sentimen vaksin COVID-19.

Parameter yang digunakan untuk performansi klasifikasi yaitu akurasi, precision dan recall. Akurasi merupakan rasio kinerja observasi yang diprediksi dengan benar dari total observasi. Akurasi dapat dihitung menggunakan Persamaan 5.

$$Akurasi = \frac{TP+TN}{TP+FP+FN+TN} \quad (5)$$

Precision merupakan rasio observasi positif yang diprediksi dengan benar dari total observasi positif yang diprediksi. Untuk mencari nilai *precision* digunakan Persamaan 6.

$$Precision = \frac{TP}{TP+FN} \quad (6)$$

Sedangkan *recall*, bisa disebut *sensitivity* adalah rasio pengamatan positif yang diprediksi dengan benar untuk semua pengamatan di kelas yang sebenarnya. Nilai recall dapat dihitung menggunakan Persamaan 7.

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

3. Hasil dan Pembahasan

Algoritma klasifikasi yang dikaji pada penelitian ini yaitu RNN dan Naïve Bayes dengan menerapkan teknik TF-IDF menggunakan dataset *tweet* vaksin COVID-19. Pengujian dilakukan dengan membagi komposisi data latih sebesar 80% dan data uji sebesar 20% dari dataset. Hasil klasifikasi kedua metode tersebut dibandingkan untuk memperoleh metode yang tepat dalam mengklasifikasikan sentimen vaksin COVID-19.

| Layer (type) | Output Shape | Param # |
|-----------------------------|------------------|---------|
| Embedding_layer (Embedding) | (None, 100, 250) | 1250000 |
| LSTM_layer (LSTM) | (None, 80) | 105920 |
| Output_layer (Dense) | (None, 3) | 243 |
| Total params: 1,356,163 | | |
| Trainable params: 1,356,163 | | |
| Non-trainable params: 0 | | |

Gambar 6. Parameter RNN

Pada Gambar 6. menunjukkan bahwa *hidden layer* yang digunakan sebesar 100 dengan *embedding size* 250. *Layer* yang digunakan sebanyak 80 dengan *dense* sebesar 3. Parameter untuk mengetahui *loss* yaitu *binary crossentropy* dengan fungsi aktivasi *softmax* dan *optimizer adam*.

Tabel 6. Hasil Akurasi RNN

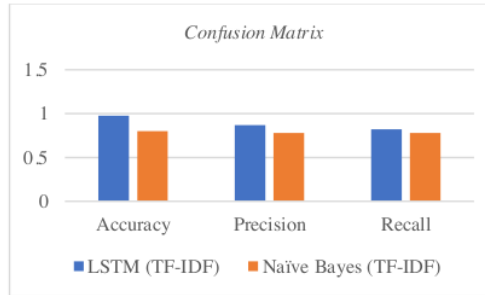
| Epoch | 10 | 20 | 30 | 40 | 50 |
|--------------|--------|--------|--------|--------|--------|
| Data Uji (%) | | | | | |
| 20 | 0.8311 | 0.9770 | 0.8122 | 0.7971 | 0.8103 |
| 30 | 0.8132 | 0.8050 | 0.7893 | 0.8138 | 0.8012 |
| 40 | 0.8165 | 0.8150 | 0.8240 | 0.8141 | 0.4207 |

Pada Tabel 6. dapat dilihat pengujian data uji secara berturut-turut 20%, 30% dan 40% menggunakan metode RNN (TF-IDF). Hasil terbaik yang didapatkan pada model RNN adalah sebesar 97,77% dengan data uji sebesar 20%. Hal tersebut juga diperlakukan sama pada metode Naïve Bayes (TF-IDF) dan mendapatkan hasil terbaik dengan komposisi data latih sebesar 80% dengan data uji sebesar 20% seperti yang terlihat pada Tabel 7.

Tabel 7. Hasil Akurasi Naïve Bayes

| Test Size | Akurasi |
|-----------|---------|
| 0,2 | 0.79 |
| 0,3 | 0.80 |
| 0,4 | 0.79 |

Perbandingan *Confusion Matrix* metode RNN dan Naïve Bayes menggunakan TF-IDF ditunjukkan pada Tabel 5. Kajian hasil dari kedua metode tersebut dievaluasi berdasarkan hasil akurasi, *precision* dan *recall*. RNN (TF-IDF) mampu mengklasifikasikan dengan nilai akurasi 0,977, nilai *precision* 0,87 dan nilai *recall* 0,82. Klasifikasi metode *naïve bayes* menghasilkan nilai akurasi sebesar 0,80, nilai *precision* 0,78 dan nilai *recall* 0,78.



Gambar 7. Grafik Hasil Confusion Matrix

Perbandingan akurasi antar metode klasifikasi ditunjukkan pada Gambar 7. Metode SVM dan Naïve Bayes pada [24] menghasilkan akurasi sebesar 82,48% dan 76,56%. Penelitian lain dengan metode Naïve Bayes, SVM dan KNN mendapatkan hasil akurasi sebesar 75,58%, 63,99% dan 73,34%[25]. Selain itu, akurasi metode Naïve Bayes, SVM dan *Random Forest* pada [26] yaitu 74,2%, 81,2% dan 72,5%. Hasil akurasi RNN [27]–[29] berturut-turut sebesar 91,9%, 85% dan 95%. Metode klasifikasi yang digunakan pada penelitian ini menunjukkan hasil akurasi yang relatif tidak jauh berbeda. Dari hasil simulasi, implementasi metode RNN dan Naïve Bayes menggunakan teknik TF-IDF untuk sentimen analisis vaksin COVID-19 mampu melakukan perbaikan yang ditunjukkan dengan nilai akurasi yang lebih tinggi. Nilai akurasi metode RNN (TF-IDF) 97,77% dan Naïve Bayes (TF-IDF) 80%. Jika dibandingkan dengan [24]–[29], metode RNN(TF-IDF) memiliki nilai akurasi yang tinggi yaitu sebesar 97,77%. Perbandingan akurasi antar metode ditunjukkan pada Tabel 8.

Tabel 8. Perbandingan Akurasi Antar Metode

| Metode | Akurasi |
|----------------------|---------|
| SVM[24] | 82.48% |
| Naïve Bayes[24] | 76.56% |
| Naïve Bayes[25] | 75.58% |
| SVM[25] | 63.99% |
| KNN[25] | 73.34% |
| Naïve Bayes[26] | 74.2% |
| SVM[26] | 81.2% |
| Random Forest[26] | 72.5% |
| RNN[27] | 91.9% |
| RNN[28] | 85% |
| RNN[29] | 95% |
| Naïve Bayes (TF-IDF) | 80% |
| RNN (TF-IDF) | 97.7% |

4. Kesimpulan

Penelitian ini berhasil dilakukan dengan membandingkan kinerja beberapa metode RNN dan Naïve Bayes menggunakan teknik pembobotan TF-IDF. Dataset pada penelitian ini diperoleh dari 5000 *tweet* vaksin COVID-19 dengan pembagian 3800 *tweet* sentimen positif, 800 *tweet* sentimen negatif dan 400 *tweet* sentimen netral. Metode RNN (TF-IDF)

menunjukkan hasil akurasi yang paling baik yaitu sebesar 97,77% dibandingkan dengan Naïve Bayes (TF-IDF) dengan nilai akurasi sebesar 80%.

Ucapan Terimakasih

Tim peneliti mengucapkan terima kasih dan penghargaan kepada Direktorat Penelitian dan Pengabdian Kepada Masyarakat (DPPM) Universitas Muhammadiyah Malang atas dukungan terselenggaranya penelitian ini melalui skema Penelitian dan Pengembangan Ipteks (P2I) Tahun 2021.

Daftar Rujukan

- [1] G. Peretto, S. Sala, and A. L. P. Caforio, "The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak – an update on the status," *Eur. Heart J.*, vol. 41, no. 22, pp. 2124–2125, 2020, doi: 10.1093/eurheartj/ehaa396.
- [2] L. Y. C. Wong and J. Burkell, "Motivations for sharing news on social media," *ACM Int. Conf. Proceeding Ser.*, vol. Part F1296, 2017, doi: 10.1145/3097286.3097343.
- [3] Y. Shi *et al.*, "Knowledge and attitudes of medical staff in Chinese psychiatric hospitals regarding COVID-19," *Brain, Behav. Immun. - Heal.*, vol. 4, p. 100064, 2020, doi: 10.1016/j.bbih.2020.100064.
- [4] M. A. Fauzi, "Random forest approach fo sentiment analysis in Indonesian language," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 12, no. 1, pp. 46–50, 2018, doi: 10.11591/ijeecs.v12.i1.pp46-50.
- [5] W. K. Sari, D. P. Rini, and R. F. Malik, "Text Classification Using Long Short-Term Memory With GloVe Features," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 5, no. 2, p. 85, 2020, doi: 10.26555/jiteki.v5i2.15021.
- [6] W. K. Sari, D. P. Rini, R. F. Malik, and I. S. B. Azhar, "Klasifikasi Teks Multilabel pada Artikel Berita Menggunakan Long Short-Term Memory dengan Word2Vec," *Resti*, vol. 1, no. 10, pp. 276–285, 2017.
- [7] M. Hung *et al.*, "Social network analysis of COVID-19 sentiments: Application of artificial intelligence," *J. Med. Internet Res.*, vol. 22, no. 8, 2020, doi: 10.2196/22590.
- [8] T. Hendrawati and C. P. Yanti, "Analysis of Twitter Users Sentiment against the Covid-19 Outbreak Using the Backpropagation Method with Adam Optimization," *J. Electr. Electron. Informatics*, vol. 5, no. 1, p. 1, 2021, doi: 10.24843/jeei.2021.v05.i01.p01.
- [9] A. K. Fauziyah, "Analisis Sentimen Pandemi Covid19 Pada Streaming Twitter Dengan Text Mining Python," *J. Ilm. SINUS*, vol. 18, no. 2, p. 31, 2020, doi: 10.30646/sinus.v18i2.491.
- [10] N. Chintalapudi, G. Battineni, and F. Amenta, "Sentimental analysis of COVID-19 tweets using deep learning models," *Infect. Dis. Rep.*, vol. 13, no. 2, pp. 329–339, 2021, doi: 10.3390/IDR13020032.
- [11] T. Vijay, A. Chawla, B. Dhanka, and P. Karmakar, "Sentiment Analysis on COVID-19 Twitter Data," *2020 5th IEEE Int. Conf. Recent Adv. Innov. Eng. ICRAIE 2020 - Proceeding*, vol. 2020, no. November 2019, 2020, doi: 10.1109/ICRAIE51050.2020.9358301.
- [12] J. V. Lazarus *et al.*, "A global survey of potential acceptance of a COVID-19 vaccine," *Nat. Med.*, vol. 27, no. 2, pp. 225–228, 2021, doi: 10.1038/s41591-020-1124-9.
- [13] K. M. Bubar *et al.*, "Model-informed COVID-19 vaccine prioritization strategies by age and serostatus," *Science (80-.)*, vol. 371, no. 6532, pp. 916–921, 2021, doi: 10.1126/science.abe6959.
- [14] F. F. Rachman and S. Pramana, "Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter," *Heal. Inf. Manag. J.*, vol. 8, no. 2, pp. 100–109, 2020, [Online]. Available: <https://inohim.esaunggul.ac.id/index.php/INO/article/view/223/175>.
- [15] C.- Pandemic, B. Laurensz, and E. Sedyono, "Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19 (Analysis of Public Sentiment on Vaccination in Efforts to Overcome the," vol. 10, no. 2, pp. 118–123, 2021.
- [16] D. A. Nurdeni, I. Budi, and A. B. Santoso, "Sentiment Analysis on Covid19 Vaccines in Indonesia: From the Perspective of Sinovac and Pfizer," *3rd 2021 East Indones. Conf. Comput. Inf. Technol. EIConCIT 2021*, no. April, pp. 122–127, 2021, doi: 10.1109/EIConCIT50028.2021.9431852.
- [17] F. W. Ramadhan, H. T. Sukmana, L. K. Oh, and ..., "Analysis Of Warganet Comments On It Services In Mandiri Bank Using K-Nearest Neighbor (K-Nn) Algorithm Based On Ism Criteria," *ADI Journal on Recent academia.edu*, 2019, [Online]. Available: https://www.academia.edu/download/60832373/Paper_3_Fix_bgt_say_20191008-60662-1rud270.pdf.
- [18] K. Setiawan, B. Rahmatullah, and ..., "Komparasi Metode Naive Bayes Dan Support Vector Machine Menggunakan Particle Swarm Optimization Untuk Analisis Sentimen" *J.*, 2020, [Online]. Available: <http://journal.stmikjayakarta.ac.id/index.php/jisamar/article/view/250>.
- [19] M. Cindo, D. P. Rini, and E. Ermatita, "Literatur Review: Metode Klasifikasi Pada Sentimen Analisis," *Semin. Nas. Teknol.*, 2019, [Online]. Available: <http://seminar-id.com/prosiding/index.php/sainteks/article/view/124>.
- [20] A. F. Anees, A. Shaikh, A. Shaikh, and S. Shaikh, "Survey Paper on Sentiment Analysis : Techniques and Challenges," *EasyChair*, pp. 2516–2314, 2020.
- [21] N. Yadav, O. Kudale, A. Rao, S. Gupta, and A. Shitole, "Twitter Sentiment Analysis Using Supervised Machine Learning," *Lect. Notes Data Eng. Commun. Technol.*, vol. 57, no. April 2020, pp. 631–642, 2021, doi: 10.1007/978-981-15-9509-7_51.
- [22] Imamah and F. H. Rachman, "Twitter sentiment analysis of Covid-19 using term weighting TF-IDF and logistic regression," *Proceeding - 6th Inf. Technol. Int. Semin. ITIS 2020*, pp. 238–242, 2020, doi: 10.1109/ITIS50118.2020.9320958.
- [23] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, and A. E. Hassanien, "Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media," *Appl. Soft Comput. J.*, vol. 97, p. 106754, 2020, doi: 10.1016/j.asoc.2020.106754.
- [24] A. M. Rahat, A. Kahir, and A. K. M. Masum, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset," *Proc. 2019 8th Int. Conf. Syst. Model. Adv. Res. Trends. SMART 2019*, pp. 266–270, 2020, doi: 10.1109/SMART46866.2019.9117512.
- [25] M. Wongkar and A. Angdresy, "Sentiment analysis using Naive Bayes Algorithm of the data crawler: Twitter," *2019 Fourth Int.*, 2019, [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8985884>.
- [26] G. A. Ruz, P. A. Henriquez, and A. Mascareño, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers," *Futur. Gener. Comput. Syst.*, vol. 106, pp. 92–104, 2020, doi: 10.1016/j.future.2020.01.005.
- [27] L. Kumiasari and A. Setyanto, "Sentiment Analysis using Recurrent Neural Network," *J. Phys. Conf. Ser.*, vol. 1471, no. 1, 2020, doi: 10.1088/1742-6596/1471/1/012018.
- [28] B. N. Saha, A. Senapati, and A. Mahajan, "LSTM based Deep RNN Architecture for Election Sentiment Analysis from Bengali Newspaper," *2020 Int. Conf. Comput. Perform. Eval. ComPE 2020*, pp. 564–569, 2020, doi: 10.1109/ComPE49325.2020.9200062.
- [29] M. F. Wahid, M. J. Hasan, and M. S. Alom, "Cricket Sentiment Analysis from Bangla Text Using Recurrent Neural Network with Long Short Term Memory Model," *2019 Int. Conf. Bangla Speech Lang. Process. ICBSLP 2019*, no. September, pp. 27–28, 2019, doi: 10.1109/ICBSLP47725.2019.201500.

DOI: <https://doi.org/10.29207/resti.v5i4.3308>

Creative Commons Attribution 4.0 International License (CC BY 4.0)

- [30] M. Atikur Rahman and E. K. Dey, "Datasets for aspect-based sentiment analysis in bangla and its Baseline evaluation," *Data*, vol. 3, no. 2, 2018, doi: 10.3390/data3020015.
- [31] M. Savargiv, B. Masoumi, and M. R. Keyvanpour, "A new random forest algorithm based on learning automata," *Comput. Intell. Neurosci.*, vol. 2021, 2021, doi: 10.1155/2021/5572781.
- [32] A. M. Pravina, I. Cholissodin, and P. P. Adikara, "Analisis Sentimen Tentang Opini Maskapai Penerbangan pada Dokumen Twitter Menggunakan Algoritme Support Vector Machine (SVM)," ... *Teknol. Inf. dan Ilmu ...*, 2019, [Online]. Available: <http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/4793>.
- [33] M. Allahyari *et al.*, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," 2017, [Online]. Available: <http://arxiv.org/abs/1707.02919>.
- [34] S. Ghosh and M. S. Desarkar, "Class Specific TF-IDF Boosting for Short-text Classification," pp. 1629–1637, 2018, doi: 10.1145/3184558.3191621.
- [35] A. Z. Amrullah, A. S. Anas, and ..., "Analisis Sentimen Movie Review Menggunakan Naive Bayes Classifier Dengan Seleksi Fitur Chi Square," *J. ...*, 2020, [Online]. Available: <https://journal.universitasbumigora.ac.id/index.php/bite/article/view/804>.
- [36] M. Wongkar and A. Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter," *Proc. 2019 4th Int. Conf. Informatics Comput. ICIC 2019*, pp. 1–5, 2019, doi: 10.1109/ICIC47613.2019.8985884.

Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naïve Bayes

ORIGINALITY REPORT

21%
SIMILARITY INDEX

18%
INTERNET SOURCES

9%
PUBLICATIONS

8%
STUDENT PAPERS

PRIMARY SOURCES

- 1** Nufia Alfi Rohyana, Aris Wahyu Murdiyanto, Kharisma. "Analisis Sentimen Di Media Sosial Twitter Dengan Studi Kasus Vaksinasi Covid-19", *INDONESIAN JOURNAL ON DATA SCIENCE*, 2023
Publication **2%**
- 2** Submitted to Higher Education Commission Pakistan
Student Paper **2%**
- 3** ejournal.poltektegal.ac.id
Internet Source **2%**
- 4** www.scribd.com
Internet Source **2%**
- 5** eprints.amikom.ac.id
Internet Source **1%**
- 6** Submitted to Universitas International Batam
Student Paper **1%**
- 7** senafti.budiluhur.ac.id
Internet Source **1%**

| | | |
|----|--|-----|
| 8 | 123dok.com Internet Source | 1 % |
| 9 | jurnal.untan.ac.id Internet Source | 1 % |
| 10 | openlibrarypublications.telkomuniversity.ac.id Internet Source | 1 % |
| 11 | journal.unsika.ac.id Internet Source | 1 % |
| 12 | jtiulm.ti.ft.ulm.ac.id Internet Source | 1 % |
| 13 | Submitted to Universitas Singaperbangsa Karawang Student Paper | 1 % |
| 14 | repo.undiksha.ac.id Internet Source | 1 % |
| 15 | jtsiskom.undip.ac.id Internet Source | 1 % |
| 16 | Haris Ahmad Gozali, Mochamad Alfian Rosid, Sumarno. "Classification of Student Complaints with the Naive Bayes and Literature Methods", JOINCS (Journal of Informatics, Network, and Computer Science), 2020 Publication | 1 % |
| 17 | ejournal.itn.ac.id Internet Source | |

1 %

18

journal.universitaspahlawan.ac.id
Internet Source

1 %

19

jurnal.unsil.ac.id
Internet Source

1 %

Exclude quotes On

Exclude matches < 1%

Exclude bibliography On



Digital Receipt


This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author: Artikel 6
Assignment title: Lailis Syfaah 4
Submission title: Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Re...
File name: OVID-19_Menggunakan_Recurrent_Neural_Network_dan_Na...
File size: 392.23K
Page count: 7
Word count: 4,659
Character count: 26,397
Submission date: 14-Mar-2024 10:12AM (UTC+0700)
Submission ID: 2319966654

Terakreditasi SINTA Peringkat 2
Serial Keptuhan Ditjen Pengantian Riset dan Pengembangan Riset Dikti No. 100/KP/2019

Terbit online pada laman web jurnal: <http://jurnal.iainior.id>

**JURNAL RESTI**
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 5 No. 2 (2021) 802 - 808 ISSN Media Elektronik: 2580-0760

Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naive Bayes

Merinda Lestandy¹, Abdurrahim Abdurrahim², Lailis Syafa'ah³
¹Program Studi D3 Teknologi Elektronika, Fakultas Teknik, Universitas Muhammadiyah Malang
²Program Studi Teknik Elektro, Fakultas Teknik, Universitas Muhammadiyah Malang
³merinda.lestandy@umm.ac.id, abdurrahim@webmail.umm.ac.id, lailis@umm.ac.id

Abstrak
COVID-19 has become a global pandemic including Indonesia, so the government is taking vaccinations as a preventive measure. The public's response to this continues to appear on social media platforms, one of which is Twitter. Tweets about the COVID-19 vaccine have generated various kinds of positive and negative opinions in the community. Therefore, it is very important to detect and filter it to prevent the spread of incorrect information. Sentiment analysis is a method used to determine the content of a dataset in the form of negative, positive or neutral text. The dataset in this study was obtained from 5000 COVID-19 vaccine tweets with the distribution of 3800 positive sentiment tweets, 800 negative sentiment tweets and 400 neutral sentiment tweets. The dataset obtained is then pre-processed data to optimize data processing. There are 4 stages of pre-processing, including remove punctuation, case folding, stemming and lemmatizing. This study examines the performance of RNN and Naive Bayes by adding the TF-IDF (Term Frequency-Inverse Document Frequency) technique which aims to give weight to the word relationship terms of a document. The test results show that RNN (TF-IDF) has a greater accuracy of 97.77% compared to Naive Bayes (TF-IDF) of 80%.

Keywords: Sentiment Analysis, Vaccine COVID-19, TF-IDF, RNN, Naive Bayes

Abstrak
COVID-19 telah menjadi pandemi dunia termasuk Indonesia, sehingga pemerintah mengambil tindakan vaksinasi sebagai upaya pencegahan. Tanggapan masyarakat mengenai hal tersebut terus muncul di platform media sosial salah satunya twitter. Tweet mengenai vaksin COVID-19 menimbulkan berbagai macam opini positif dan negatif di masyarakat. Oleh karena itu, sangat penting untuk mendeteksi dan menyaringnya agar tidak terjadi penyebaran informasi yang tidak benar. Analisis sentimen merupakan suatu metode yang digunakan untuk menentukan isi dari suatu dataset dalam bentuk teks yang bersifat negatif, positif atau netral. Dataset pada penelitian ini diperoleh dari 5000 tweet vaksin COVID-19 dengan pembagian 3800 tweet sentimen positif, 800 tweet sentimen negatif dan 400 tweet sentimen netral. Dataset yang didapat kemudian dilakukan pre-processing data untuk mengoptimalkan pengolahan data. Terdapat 4 tahapan pre-processing antara lain remove punctuation, case folding, stemming dan lemmatizing. Penelitian ini mengkaji kinerja RNN dan Naive Bayes dengan menambahkan teknik TF-IDF (Term Frequency-Inverse Document Frequency) yang bertujuan untuk memberikan bobot pada hubungan kata (term) sebuah dokumen. Hasil pengujian menunjukkan RNN (TF-IDF) memiliki akurasi lebih besar yaitu 97,77% dibandingkan Naive Bayes (TF-IDF) sebesar 80%.

Kata kunci: Analisis Sentimen, Vaksin COVID-19, TF-IDF, RNN, Naive Bayes

1. **Pendahuluan**
akibat virus tersebut. Krisis akibat COVID-19 saat ini menciptakan situasi sosial bagi kesehatan mental masyarakat[1].
COVID-19 (Coronavirus Disease-2019) merupakan penyakit yang disebabkan oleh virus corona jenis baru yaitu Sars-Cov-2. Virus ini menyebar dengan kontak langsung melalui tetapan udara seperti batuk dan bersin, untuk pertama kalinya, dibandingkan sumber tradisional Berdasarkan data www.covid19.go.id, penyebaran virus lainnya seperti televisi, surat kabar dll [2][3]. Orang COVID-19 di Indonesia per tanggal 1 Maret 2020 cenderung percaya apa yang mereka temukan di jejaring sosial dengan 13 Agustus 2021 sebanyak 3.804.043 sosial, yang membuat mereka rentan terhadap rumor dan terkonfirmasi positif COVID-19 dan 115.096 meninggal berita palsu. Oleh karena itu, sangat penting untuk

Diterima Redaksi: 19-07-2021 | Selesai Revisi: 24-08-2021 | Diterbitkan Online: 26-08-2021
802