Klasifikasi Teks Twitter Menggunakan Algoritma Naïve Bayes untuk Analisis Sentimen Penggunaan Vaksin Covid-19

Abdul Rohim[1], **Amiq Fahmi**[2]
[1,2] Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro.

Correspondence Author: amiq.fahmi@dsn.dinus.ac.id

Abstract
Pandemi Covid-19 berdampak buruk terutama pada sektor kesehatan, ekonomi, dan pendidikan. Pemerintah Indonesia melakukan pencegahan dengan melakukan vaksin dosis ke-1 dan ke-2. Namun, dinilai masih kurang efektif untuk menghambat penyebaran virus. Selanjutnya diperkuat dengan melakukan vaksin ke 3 (booster). Tujuan dari penelitian ini untuk menganalisis sentimen pada masyarakat mengenai pelaksanaan vaksin booster. Analisis ini untuk membantu stakeholder dalam memahami sentimen masyarakat baik positif, netral, maupun negatif. Data yang digunakan sebanyak 1.122 tweet dengan menggunakan kata kunci "vaksin booster dan covid". Pada penelitian ini, kami menggunakan algoritma Naïve Bayes untuk prediksi sentimen analisis. Dataset untuk pelatihan dan pengujian sebesar 90% (1.009) dan tes 10% (113). Hasil eksperimen menghasilkan akurasi prediksi sebesar 72%, precission 68%, recall 74%, F1-score 70%, dan nilai AUC/ROC 82%. Hasil analisis sentimen "netral" sebanyak 518 (46.2 %), "positif" sebanyak 437 (38.9%), dan "negatif" sebanyak 167 (14.9%). Hasil dapat diartikan bahwa Algoritma Naïve Bayes memiliki performa klasifikasi yang baik untuk target sentimen multi-kelas.

Keyword: Analisis Sentimen, Text mining, Naïve Bayes, Vaksin Booster, Covid-19.

1. PENDAHULUAN

Pada akhir tahun 2019, ditemukan kemunculan kasus pertama yang diduga menjadi sumber penyebaran covid-19 di Kota Wuhan, Tiongkok [1]. Varian tipe virus baru ini dilabeli nama *Coronavirus Disease* (Covid-19). Virus ini menyerang sistem pernafasan manusia dan menimbulkan penyakit pernafasan kronis dan menyebar dengan cepat hingga menyebabkan pandemi diberbagai belahan dunia. Penyebaran karena interaksi langsung antar manusia, sehingga seseorang yang positif Covid-19 akan membuat lainya menjadi terpapar virus covid-19.

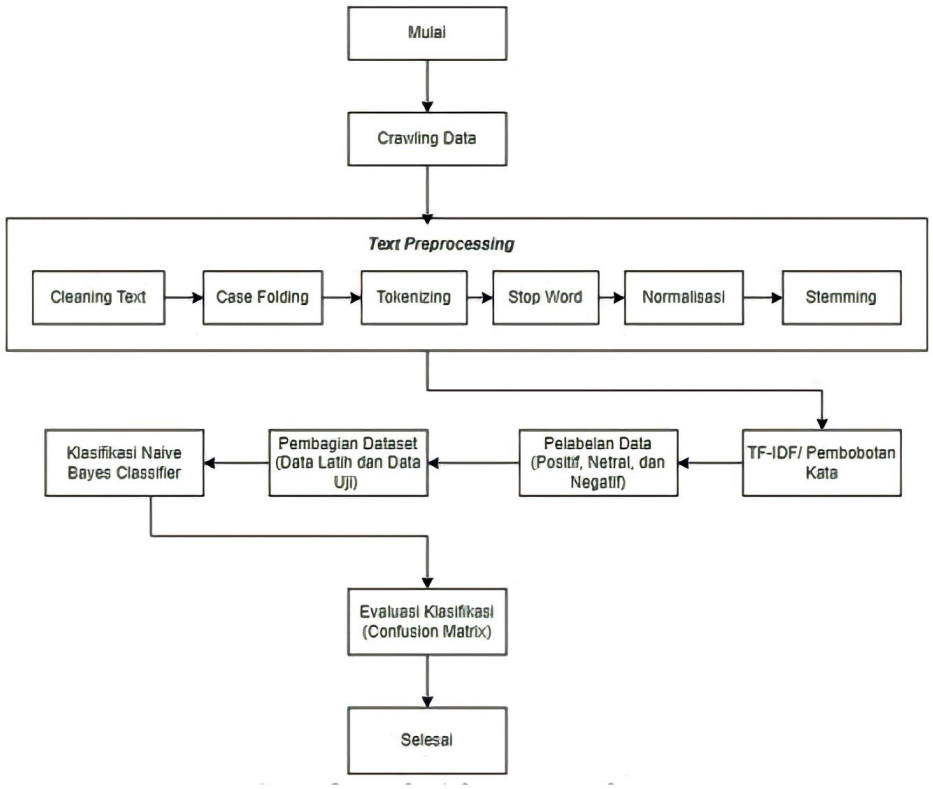
Pemerintah Indonesia terkait bencana kesehatan ini, berusaha untuk mengurangi penyebaran Covid-19 dengan melakukan menerapkan protokol kesehatan yang ketat, pembatasan sosial berskala besar (PSBB) dan vaksin COVID-19. Vaksin 1, dan 2 masih kurang efektif untuk menghambat penyebaran covid-19 di Indonesia. Akibat gelombang serangan varian *Omicron*, vaksin *booster* menjadi bahan perbincangan yang ramai di masyarakat, baik melalui media media sosial maupun berbagai respon dan pendapat yang disampaikan pada media lainnya. Satu dari beberapa media sosial yang banyak digunakan oleh masyarakat dalam memberikan aspirasi dan opini yaitu media sosial Twitter[2].

Tujuan dari penelitian adalah untuk menganalisi pandangan dan opini masyarakat yang beragam di Twitter. Penelitian ini berfokus pada analisis sentimen untuk membantu kapabilitas pemangku kepentingan melalui informasi yang bermanfaat. Dalam melakukan analisis sentimen, kami menerapkan pembelajaran mesin untuk pengolahan data, dan menggunakan model Naïve Bayes untuk mengklasifikasikannya. Diantara algoritma yang digunakan untuk penerapan analisis sentimen antara lain adalah algoritma Naïve Bayes*, K-Nearest Neighbor, Support Vector Machine*, dan lainnya [3]. Pada penelitian ini, twitter sebagai objek penelitian. Twitter secara legal memberikan hak akses pada developer untuk memperoleh dataset tweet dengan metode *crawling* data menggunakan *library* Twitter API yang tersedia pada akun Twitter Developer [4].

Penelitian terkait sentimen analisis sentimen vaksinasi Covid-l 9 menggunakn pengklasifikasi menghasilkan akurasi tinggi sebesar 85,59% dibandingkan dengan SVM yang memiliki akurasi sebesar 84,41% [5]. Peneliti lain, analisis sentimen melalui komentar di sosial media Facebook, untuk studi kasus pada akun Jasa ekspedisi barang J&T ekspress Indonesia dengan hasil akurasi tertinggi adalah 79.21% dan akurasi terendah adalah 70.3% dengan metode K-Nearest Neighbor [6]. Alasan pemilihan algoritma Naïve Bayes pada studi ini, karena Naïve Bayes cara kerjanya yang sederhana, mudah dan efisien digunakan untuk klasifikasi sentimen analisis. Hal itu ditunjukan pada penelitian terkait sebelumnya yang memperoleh hasil nilai akurasi yang lebih baik jika dibandingkan dengan algoritma lainnya [5].

2. METODE PENELITIAN

Metode yang diusulkan pada penelitian ini adalah menerapkan algoritma klasifikasi Naïve Bayes dengan objek penelitian pada media sosial Twitter. Penelitian mengangkat topik analisis sentimen terkait pelaksanaan vaksin booster Covid-l9. Metode pengumpulan data dilakukan dengan metode *crawling* data tweets di media sosial Twitter dengan twitter API yang dilakukan secara langsung dengan memakai kata kunci antara lain vaksin, *booster* dan covid. Hasil *crawling* yang dilakukan pada tanggal 23 Juni 2022 menghasilkan data sebanyak 5000 *tweets* untuk dijadikan sebagai dataset pada penelitian ini. Langkah-langkah metode penelitian secara lengkap seperti pada Gambar 1.



Gambar 1. Metode penelitian untuk analisis setimen vaksin *booster* C-19.

Alur penelitian pada Gambar 1 menjelaskan seluruh metode yang digunakan pada penelitian ini. Adapun tahapannya sebagai berikut:

1. Pengumpulan dataset dengan melakukan *crawling* data di media sosial Twitter sesuai dengan topik yang dianalisis.
2. Melakukan *Text Preprocessing* untuk mengubah data menjadi terstruktur.
3. Klasifikasi kategori dari jumlah kata yang mengandung kategori terbanyak dalam setiap kalimat. Jika jumlah kategori kata positif dan negatif seimbang, maka kalimat tersebut termasuk dalam kategori netral.
4. Pembagian data menjadi data latih dan data uji. Pengujian dilakukan sebanyak 2 kali. Pengujian pertama menggunakan data latih sebanyak 70% dan data uji sebesar 30%. Pengujian kedua dilakukan dengan *rasio* 90 % data latih dan 10% untuk data uji.
5. Jenis setiap ketentuan diperbandingkan melalui biner (benar atau salah), frekuensi atau frekuensi dokumen dan frekuensi terbalik (TF-IDF).
6. Proses klasifikasi menggunakan data latih dan uji menggunakan pengklasifikasi menggunakan Naïve Bayes.
7. Evaluasi akurasi dilakukan untuk mengetahui akurasi model yang telah diimplementasikan menggunakan *Confusion Matrix* Tabel dan Gambar disajikan *center*, seperti yang ditunjukkan di bawah ini dan harus dikutip dalam naskah.

3. HASIL DAN ANALISA

Pada bagian ini, menjelaskan hasil proses analisis sentimen masyarakat terhadap penggunaan vaksin Covid-l9 menggunakan algoritma klasifikasi Naïve Bayes pada media sosial Twitter.

3.1 *Crawling* Data

Pengumpulan data dilakukan melalui crawling data pada media sosial Twitter dengan memanfaatkan *Application Programming Integration* (API) yang diperoleh dari akun twitter developer untuk dapat mengakses data *tweet* dari akun pengguna twitter secara resmi. Selanjutnya, setelah mendapatkan kode akses API untuk *crawling* data, kami melakukan pengambilan data menggunakan bahasa pemrograman Python menggukan *library tweepy* pada *tool jupyter notebook* untuk segera melakukan proses *crawling* data twitter dengan kode akses API tersebut.

Pengambilan data sebagai dataset untuk peneliti ini berupa data atribut waktu *tweet* dibuat (waktu), identitas nomor data pengguna (*id*), akun pengguna twitter (*username*), dan data teks cuitan (*text*). Baris teks program lengkap seperti pada Gambar 2.

```
import tweepy
import csv
import pandas as pd
import string

access_token = "1360910778056855556-7ePjIK7Tk3n6n5s7svcaGQKaxsB0Zf"
access_token_secret = "IdZXvLMmJGGZQXyRHK6QFI0QNJq1ixkB9jFuBgIC2Aghj"
api_key = "I7GzBgw4JTqRUisdAQUUySKJI"
api_key_secret = "K89FmYAH6Lv2PS4yDM6X5oxjY3CIZfuTpAaSkgx0qtX17qpCHv"

auth = tweepy.OAuthHandler(api_key,api_key_secret)
auth.set_access_token(access_token,access_token_secret)
api = tweepy.API(auth,wait_on_rate_limit=True)
search_key = "vaksin, booster, covid"

csvfile = open(search_key+".csv", "a+", newline="", encoding="utf-8")
csvwriter = csv.writer(csvfile)
c = []
i = []
u = []
t = []

for tweet in tweepy.Cursor(api.search,q=search_key,count=5000,lang='id',since="2022-06-22").items(5000):
    print(tweet.created_at,tweet.id,tweet.user.name,tweet.text)
    c.append(tweet.created_at)
    i.append(tweet.id)
    u.append(tweet.user.name)
    t.append(tweet.text.encode("utf-8"))
    tweets = [tweet.created_at,tweet.id,tweet.user.name,tweet.text.encode("utf-8")]
    csvwriter.writerow(tweets)

dictTweets = {"waktu":c, "id":i,"username":u,"teks":t}
df = pd.DataFrame(dictTweets,columns=["waktu","id","username","teks"])
df
```

Gambar 2. Source code crawling dataset di Twitter

3.2 Text Preprocesing

Tahap ini merupakan sebuah proses perubahan data ke bentuk terstruktur untuk dilakukan *text mining* dengan menghapus persoalan masalah yang dapat mempengaruhi hasil pada proses data ditandai dengan teks bercetak tebal. Berikut tabel hasil dari tahapan *text preprocesing* yang dilakukan pada penelitian ini meliputi *cleansing, case folding, tokenizing, stopword removal*, *normalization*, dan *stemming* [7].

3.3 TF-IDF/ Pembobotan Kata

Proses pembobotan kata dengan menggunakan TF-IDF yaitu menghitung bobot kata dengan cara integrasi antar *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF).

Tabel 1. *Text preprocessing*

Nama Proses	**Sebelum**	**sesudah**
Cleansing	**b'G20** Indonesia Kuat cegah *Omicron* baru, Pemerintah Genjot Vaksin *Booste*r **Covid-19 https://t.co/P7ZdqIVc9U'**	Indonesia Kuat Cegah *Omicron* Barn, Pemerintah Genjot Vaksin *Booster* Covid
Case Folding	**Indonesia Kuat Cegah *Omicron* Baru, Pemerintah Genjot Vaksin *Booster***	indonesia kuat cega h *omicron* baru pemerintah genjot vaksin

	Covid	*booster* covid
Tokenizing	indonesia kuat cegah *omicron* baru pemerintah genjot vaksin *booster* covid	[‘indonesia’, ‘kuat’, ‘cegah’, ‘*omicron*’, ‘baru’, ‘pemerintah’, ‘genjot’, ‘vaksin’, ‘*booster*’, ‘covid’]
Stopword Removal	[‘indonesia’, ‘kuat’, ‘cegah’, ‘*omicron*’, **‘baru’,** ‘pemerintah’, ‘genjot’, ‘vaksin’, ‘*booster*’, ‘covid’]	[‘indonesia’, ‘kuat’, ‘cegah’, ‘*omicron*’, ‘pemerintah’, ‘genjot’, ‘vaksin’, ‘*booster*’, ‘covid’]
Normalisasi	[‘indonesia’, ‘kuat’, ‘cegah’, ‘*omicron*’, ‘pemerintah’, ‘genjot’, ‘vaksin’, ‘*booster*’, ‘covid’]	‘indonesia’, ‘kuat’, ‘cegah’, ‘*omicron*’, ‘pemerintah’, ‘genjot’, ‘vaksin’, ‘*booster*’, ‘covid’]
Stemming	[‘indonesia’, ‘kuat’, ‘cegah’, ‘*omicron*’, ‘pemerintah’, ‘genjot’, ‘vaksin’, ‘*booster*’, ‘covid’]	[‘indonesia’, ‘kuat’, ‘cegah’, ‘*omicron*’, **‘perintah’,** ‘genjot’, ‘vaksin’, ‘*booster*’, ‘covid’]

Tabel 2. Hasil proses perhitungan TF-IDF

No	Term	TF	TF-IDF
1.	Data	0.083333	0.396990
2.	Sedia	0.083333	0.389047
3.	Dukung	0.083333	0.320632
4.	efekti vitas	0.083333	0.497321
5.	Vaksin	0.083333	0.025634
6.	Covid	0.016666	0.046061
7.	Moderna	0.083333	0.279407
8.	Cegah	0.083333	0.253426
9.	Gejala	0.083333	0.281500

Tabel 2. Merupakan basil dari kode program perhitungan *Term Frequency* (TF) baris *tweet* pada *dataframe Indonesia* ke-1220 yang dapat dihitung dengan melakukan perkalian kamus dari TF dan IDF dengan nilai demi nilai yang kemudian disimpan ke *dataframe*.

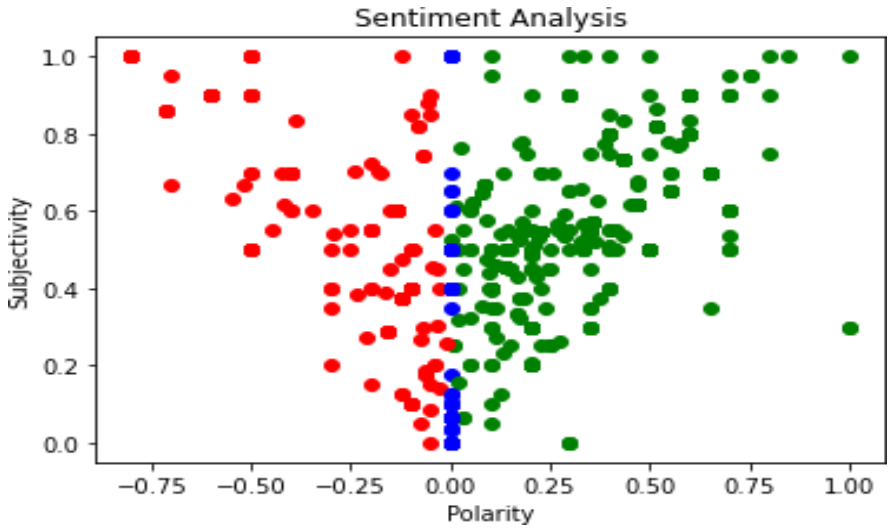
3.4 Pelabelan Data

Pelabelan dilakukan untuk melakukan analisis sentimen agar memudahkan proses klasifikasi menggunakan banyaknya dataset pada saat dilakukan pemodelan. Pelabelan ini dilakukan sesuai jumlah data *tweet* yang dimiliki, yaitu sebanyak 1122 data menggunakan modul *Library Textblob.* Modul ini digunakan untuk memproses data tekstual dikarenakan mengandung nilai *subjectivitas* dan polaritas. Gambar 3 dibawah ini Indonesia visualisasi basil dari pelabelan sentimen yang telah dilakukan pada dataset:

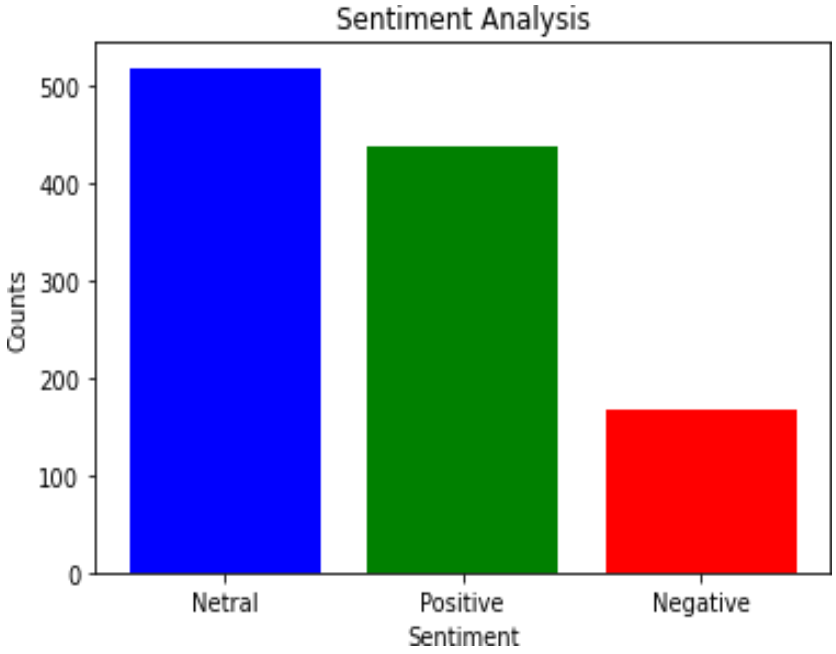
Tweet netral *scatter plot* biru mewakili distribusi *tweet* yang diklasifikasikan sebagai “netral”, *tweet positive scatter plot* hijau mewakili distribusi *tweet* yang diklasifikasikan sebagai “positive”, dan *tweet negative scatter plot* merah mewakili distribusi *tweet* yang diklasifikasikan sebagai “*negative*” menggunakan polaritas dan *subjektivitas*.

Dari pelabelan dengan jumlah data sebanyak 1.122 menunjukan bahwa label sentimen “netral” memiliki data sebanyak 518 (46.2 %), label “positif” sebanyak 437 (38.9%), dan sentimen dengan label “negatif” memiliki data sebanyak 167.

Kemudian untuk mengetahui frekuensi kata yang sering muncul pada setiap kategori label sentimen maka dilakukan visualisasi *word cloud*. *Word cloud* menampilkan visualisasi yang diperoleh dari visualisasi *tweet* berbasis kategori sentimen sehingga memungkinkan untuk melihat kata mana yang sering muncul di se tiap kata terhadap masing-masing sentimen diantaranya “netral”, “positif” dan “negatif” [8]. Ukuran *font* di *word cloud* menunjukkan seberapa sering kata tersebut muncul.



Gambar 3. Visualisasi persebaran data *score Polarity* dan *Subjectivity*

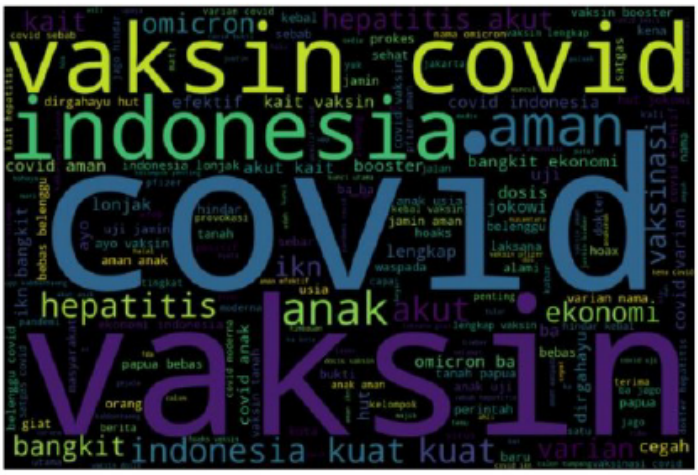


Gambar 4. Visualisasi persentase label sentimen

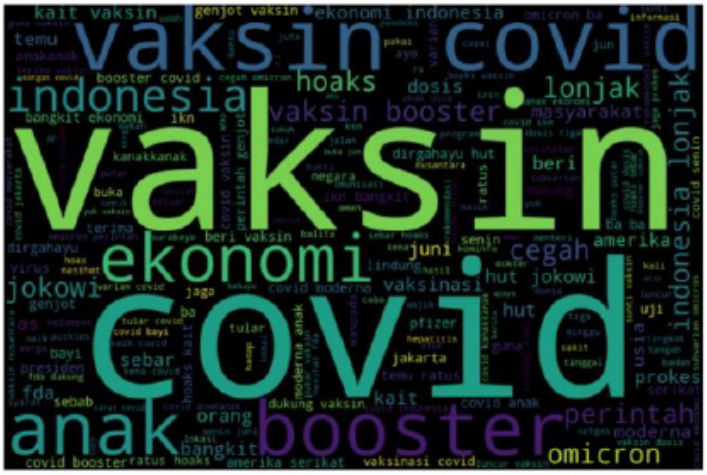
Gambar 6. Menampilkan hasil visualisasi *word cloud* dimana frekuensi kata yang sering muncul pada label sentimen "positif" diantaranya, yaitu kata covid, vaksin, Indonesia dan pada di tingkat 2 terdapat kata hepatitis, anak, ekonomi, bangkit, dan aman.

Gambar 7. Menampilkan hasil visualisasi *word cloud* dimana frekuensi kata yang sering muncul pada label sentimen netral diantaranya, yaitu kata covid, vaksin dan pada di tingkat 2 terdapat kata booster, anak, ekonomi, dan indonesia.

Gambar 8. Menampilkan hasil visualisasi *word cloud* dimana frekuensi kata yang sering muncul pada label sentimen "netral" diantaranya, yaitu kata covid, vaksin dan pada di tingkat 2 terdapat kata laksana, giat, dan himbauan.



Gambar 6 Visualisasi *Word Cloud* Sentimen Positif



Gambar 7 Visualisasi *Word Cloud* Sentimen Positif

Gambar 8 Visualisasi *Word Cloud* Sentimen Positif

3.5 Perbandingan Evaluasi Uji Performa

Tabel 3 meruapakan hasil perbandingan evaluasi performa pengujian Algoritma Naïve Bayes yang telah dilakukan sebanyak dua kali pengujian dengan rasio perbandingan yang berberda, evaluasi klasifikasi dilakukan menggunakan *confusion matrix* dan kurva ROC/AUC [9] pada setiap pengujian.

Tabel 3. Hasil uji perbandingan *splitting* dataset

No. Uji	*Ratio*	*Accuracy*	*Precission*	*Recall*	*Fl-Score*	*Roc/Auc*
1	70:30	67%	73%	69%	70%	0.81
2	90:10	72%	68%	74%	70%	0.82

Berdasarkan pengujian data yang dilakukan sebanyak 2 kali dengan nilai yang disajikan pada tabel 3., peneliti dapat mengambil kesimpulan babwa nilai akurasi tertinggi terletak pada pengujian kedua yaitu sebesar 72%, yang artinya penggunaan *split data validation* pada rasio 90: l0 memiliki performa lebib unggul dalam melakukan klasifikasi pada sentimen masyarakat terbadap penggunaan vaksin yang memiliki sebanyak 1.122 dataset diperoleb dari pengumpulan data menggunakan metode *crawling* data pada media sosial Twitter.

3.6 Analisa Hasil

Mengacu terbadap pada basil pengujian yang sudab dilakukan sebelumnya, maka diperoleh analisis hasil sebagai berikut:

1. Berdasarkan proses evaluasi performa klasifikasi Naïve Bayes [11] yang mendapatkan keputusan babwa penggunaan *split validation* pada rasio 90:10 mendapatkan nilai akurasi sebesar 72% dan nilai AUC sebesar 82%, artinya penggunaan algoritma Naïve Bayes termasuk ke dalam kategori *good classification* karena memiliki performa lebih unggul dalam melakukan klasifikasi sentimen masyarakat terbadap penggunaan vaksin *booster* Covid-19
2. Jumlah data sebanyak 1.122 menunjukan bahwa data teks yang dilakukan pelabelan menunjukan basil sentimen netral memiliki data sebanyak 518 (46.2%), label positif 437 (38.9%), dan sentiment dengan label negatif memiliki data sebanyak 167 (14.9%).
3. Kesalahan pengujian akurasi dalam penelitian ini dipengaruhi oleh kesalaban uji sistematis multi kelas, dimana data termasuk pada kelas “netral” dan sebagian data pada kelas “positif”. Begitu pula beberapa *instance* data diberi label dan memiliki makna “netral”. Terdapat beberapa kata dalam data, tetapi sistem mengklasifikasikan emosi ke dalam kategori “positif” atau “negatif”, yang bersifat makian atau sanjungan, tetapi mesin mengklasifikasikan sentimen bukan dalam kategorinya.

4. KESIMPULAN

Berdasarkan hasil analisis dan pengujian yang telah dilakukan dapat dinyatakan bahwa hasil klasifikasi menggunakan dataset Twitter dan Naïve Bayes lebih unggul dalam studi kasus sentimen vaksin *booster* Covid-19. Pada analisis sentimen menggunakan dataset sebanyak 1.122 menunjukan bahwa data teks yang dilakukan pelabelan yang terbagi ke dalam 3 kategori positif, netral dan negatif. Hasil analisis sentimen netral memiliki data sebanyak 518 (46.2%), label positif 437 (38.9%), dan sentimen dengan label negatif memiliki data sebanyak 167 (14.9%). Pemilihan metode Naïve Bayes pada proses pengujian dengan *split validation* rasio 90:10 mendapatkan nilai akurasi sebesar 72%, *precision* 68%, *recall* 74 %, *F1-score* 70%, dan nilai AUC/ROC sebesar 82%. Dengan demikian penggunaan algoritma Naïve Bayes termasuk ke dalam kategori *good classification* pada sentimen masyarakat terhadap penggunaan vaksin *booster* Covid-19.

DAFTAR PUSTAKA

[1] I. P. Sari and S. Sriwidodo, "Perkembangan Teknologi Terkini dalam Mempercepat Produksi Vaksin COVID-19," Maj. Farmasetika, vol. 5, no. 5, p. 204, 2020, doi: 10.24198/mfarmasetika. v5i5 .28082.

[2] Ratino, N. Hafidz, S. Anggraeni, and W. Gata, "Sentimen Analisis Informasi Covid-19 menggunakan *Support Vector Machine dan* Naive Bayes," J. JUPITER, vol. 12, no. 2, pp. 1-11, 2020.

[3] R. Yasmin, "Covid-19 Menggunakan Metode Naive Bayes *Classifier* Pada Media Sosial Twitter Covid-19 Menggunakan Metode Naive Bayes," 2021.

[4] AM. Rizki, "Analisis Sentimen Pada Masyarakat Tentang Penggunaan Vaksin Sinovac Menggunakan Naive Bayes *Classifier* Pada Media Sosial Twitter," STMIK Riau, vol. 365, p. 63, 2022, [Online]. Available: http://repository.sar.ac.id/id/eprint/365/1/M. ZIKRI ALFA ROBBY (1710031802087). pdf.

[5] B. Laurensz and Eko Sediyono, "Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19," J. Nas. Tek. Elektra dan Teknol. Inf, vol. 10, no. 2, pp. 118-123, 2021, doi: 10.22146/jnteti.v10i2.1421.

[6] A. Salam, J. Zeniarja, and R. S. U. Khasanah, "Analisis Sentimen Data Komentar Sosial Media Facebook Dengan K-Nearest Neighbor (Studi Kasus Pada Akun Jasa Ekspedisi Barang J&T Ekpress Indonesia)," Pros. SINTAK, pp. 480-486, 2018.

[7] Mussalimun, E. H. Khasby, G. I. Dzikrillah, and Muljono, “Comparison of K- N earest Neighbor (K -NN) and Naïve Bayes *Algorithm for Sentiment Analysis on Google Play Store Textual Reviews*,” in 2021 8th *International Conference on Information Technology, Computer and Electrical Engineering* (ICITACEE), Sep. 2021, pp. 180–184. doi: 10.1109/ICITACEE53184.2021.9617217.

[8] M. A. Hearst, E. Pedersen, L. Patil, E. Lee, P. Laskowski, and S. Franconeri, "*An Evaluation of Semantically Grouped Word Cloud Designs*," IEEE Trans. Vis. Comput. Graph., vol. 26, no. 9, pp. 2748–2761, Sep. 2020, doi: 10.1109/TVCG.2019.2904683.

[9] M. Ohsaki, P. Wang, K. Matsuda, S. Katagiri, H. Watanabe, and A. Ralescu, "*Confusion-Matrix-Based Kernel Logistic Regression for Imbalanced Data Classification*," IEEE Trans. Knowl. Data Eng., vol. 29, no. 9, pp. 1806–1819, Sep. 2017, doi: 10.1109/TKDE.2017.2682249

[10] A. Fahmi, E. Sugiarto, A. Winarno, S. Sumpeno, and M. H. Purnomo, "*Waqf Lands Assets Classification Based On Productive Value For Business Development Using* Naïve Bayes," in *2018 International Seminar on Research of Information Technology and Intelligent Systems* (ISRITI), Nov. 2018, pp. 622–626. doi: 10.1109/ISRITI.2018.8864489.