JIfoTech (Journal of Information Technology)
Vol. 05, No. 02, September 2025

Analisis Sentimen Publik Terhadap Vaksin Astrazeneca di Indonesia Menggunakan Multinomial Naive Bayes

Nur Qodariyah Fitriyah[1], Guruh Wijaya[2], M. Fadhil Al Hikam Tirta Bayu Aj[3], Hardian Oktavianto[4]
[1234]Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Jember
[1]nurfitriyah@unmuhjember.ac.id, [2]guruh.wijaya@unmuhjember.ac.id, [3]ajfadhil7@gmail.com, [4]hardian@unmuhjember.ac.id

***Abstract*— The COVID-19 virus, which first emerged in Wuhan in 2019, has spread globally, as one of the effect, various vaccination efforts despite ongoing debates over the long-term effectiveness of the vaccines. The Indonesian government sought to accelerate its vaccination program by securing millions of vaccine doses; however, public responses varied, with concerns primarily related to safety, effectiveness, and religious considerations. This study aims to analyze public opinion regarding the COVID-19 vaccine, specifically the AstraZeneca vaccine, by examining sentiment data from Twitter. Data were collected using the Twitter API and underwent preprocessing steps such as tokenization, stopword removal, and stemming. The processed data were then converted into numerical form using the TF-IDF method before being classified using the Multinomial Naive Bayes algorithm. The model's performance was evaluated using metrics such as accuracy, precision, recall, and F1-score to identify positive, negative, or neutral sentiments. The results show that the Multinomial Naive Bayes model effectively classified public sentiment toward the side effects of the AstraZeneca vaccine, resulting an accuracy of 86.91%, precision of 89.26%, and recall of 85.26%. These results indicate that the model is highly effective in identifying positive sentiments with a high level of accuracy.**

***Intisari*—Virus COVID-19 yang pertama kali muncul di Wuhan pada tahun 2019 telah menyebar ke seluruh dunia, mendorong berbagai upaya vaksinasi meskipun efektivitas jangka panjang vaksin masih menjadi perdebatan. Pemerintah Indonesia berupaya mempercepat program vaksinasi dengan mengamankan jutaan dosis vaksin, namun respons masyarakat beragam, terutama terkait kekhawatiran terhadap aspek keamanan, efektivitas, dan pertimbangan keagamaan. Penelitian ini bertujuan untuk menganalisis opini publik mengenai vaksin COVID-19, khususnya vaksin AstraZeneca, melalui data sentimen di platform Twitter. Data dikumpulkan menggunakan Twitter API, kemudian melalui proses pembersihan dan praproses, termasuk tokenisasi, penghapusan stopword, dan stemming. Selanjutnya, data dikonversi ke dalam bentuk numerik menggunakan metode TF-IDF sebelum diklasifikasikan menggunakan algoritma Multinomial Naive Bayes. Kinerja model dievaluasi berdasarkan metrik akurasi, presisi, recall, dan F1-score untuk mengidentifikasi sentimen positif, negatif, atau netral. Hasil penelitian menunjukkan bahwa model Multinomial Naive Bayes mampu mengklasifikasikan sentimen publik terhadap efek samping vaksin AstraZeneca dengan akurasi 86,91%, presisi 89,26%, dan recall 85,26%. Temuan ini menunjukkan bahwa model tersebut efektif dalam mengidentifikasi sentimen positif dengan tingkat ketepatan yang tinggi.**

***Kata Kunci*— AstraZeneca, COVID-19 vaccine, multinomial naive bayes, public sentiment, twitter analysis**

I. PENDAHULUAN

Diduga virus corona berasal dari kelelawar tahun lalu di Wuhan pada tahun 2019. Dengan penyebaran virus melalui penularan dari manusia ke manusia, virus ini mencapai manusia melalui penularan tersebut [1]. Virus ini telah menyebar ke 188 negara dan 25 wilayah di seluruh dunia sejak awal mulanya pada bulan November 2019, yang mendorong upaya-upaya rumit oleh WHO dan pemerintah untuk mengendalikan wabah ini, terutama karena sifat virus ini yang sangat menular. Hingga Januari 2021, dengan 2.066.175 kematian, 95.612.831 kasus yang dikonfirmasi telah dilaporkan secara global [2]. Laporan Satgas Covid-19 Indonesia menunjukkan bahwa saat ini terdapat 27.203 kematian, dengan kasus yang dilaporkan melebihi 951.651, termasuk yang tertinggi di Asia [3].

Teknik vaksinasi yang paling efektif belum dapat diketahui karena COVID-19 masih baru bagi manusia dan esensi respons imun defensif masih kurang diketahui secara valid. Oleh karena itu kebutuhan merancang variasi dan metode vaksin secara bersamaan menjadi penting [4]. Para ahli di seluruh dunia telah berlomba-lomba untuk memproduksi vaksin COVID-19 sejak epidemi dimulai, dengan setidaknya 166 kandidat vaksin kini dalam tahap produksi praklinis dan klinis [5]. Paradigma pengembangan vaksin pandemi baru telah disarankan untuk mengatasi kebutuhan mendesak akan vaksin, dengan mempersingkat periode pengembangan dari 10-15 tahun menjadi 1-2 tahun [6]. Namun, masih belum ada klarifikasi mengenai apa yang dapat menjadi strategi vaksin COVID-19 yang aman dan efisien secara imunologis, bagaimana mengidentifikasi titik akhir yang baik dalam pengujian efikasi vaksin, dan apa yang diharapkan selama beberapa tahun ke depan dari inisiatif vaksin global. Meskipun karakteristik vaksin dapat berbeda, memberikan bukti yang jelas tentang keamanan langsung dan tidak langsung dapat membantu

merencanakan bagaimana vaksin ini dapat digunakan secara terorganisasi [7].

Presiden Republik Indonesia membentuk tim nasional untuk mempercepat produksi vaksin COVID-19. Pada tanggal 6 Oktober 2020, Presiden menandatangani dan menerbitkan Deklarasi Presiden yang menyerukan vaksin dan memulai kampanye vaksinasi untuk mengatasi pandemi COVID-19. Deklarasi tersebut menetapkan bahwa pemerintah akan mengatur penyediaan, pengiriman, dan vaksinasi vaksin COVID-19. Pemerintah Indonesia berencana untuk mengumpulkan 30 juta dosis vaksin pada akhir tahun 2020 melalui kesepakatan bilateral dengan banyak produsen vaksin dan tambahan 50 juta dosis vaksin pada awal tahun 2021.

Skema vaksinasi kemudian menjadi pro dan kontra bagi masyarakat Indonesia. Dari 112.888 responden di seluruh provinsi di Indonesia, dalam survei yang dilakukan oleh WHO bekerja sama dengan Kementerian Kesehatan RI, NITAG, dan UNICEF, sekitar 65 persen responden menyatakan kesiapan mereka untuk menerima vaksin COVID-19 yang diberikan oleh Pemerintah Indonesia, sementara sekitar 8 persen menyatakan tidak akan menerimanya. Lebih dari 27 persen responden yang tersisa menyampaikan kekhawatiran mengenai niat pemerintah untuk memberikan vaksin COVID-19. Sekitar 30 persen responden menyatakan bahwa mereka atau orang-orang terdekat mereka, seperti anggota keluarga, rekan kerja, atau tetangga, telah tertular infeksi COVID-19, dan mereka ditemukan lebih cenderung mempertimbangkan vaksin. Responden mengajukan pertanyaan substansial tentang keamanan dan khasiat vaksin, menyatakan kurangnya kepercayaan pada vaksin, dan mengajukan keraguan tentang jenis vaksin haram-halal. Penjelasan yang paling umum untuk tidak divaksinasi adalah kekhawatiran keamanan vaksin 30%; keraguan terhadap kemanjuran vaksin 22%; hilangnya kepercayaan terhadap vaksinasi 13%; kekhawatiran terhadap efek samping seperti demam dan rasa tidak nyaman 12%; dan keyakinan agama 8%.

Urgensi penelitian ini terletak pada pentingnya memahami persepsi publik terhadap vaksin AstraZeneca di Indonesia pada tahun 2025, ketika isu kepercayaan terhadap program imunisasi masih menjadi tantangan pasca-pandemi COVID-19. Meskipun vaksinasi massal telah berlangsung sejak 2021, perbedaan opini masyarakat yang terekspresi di media sosial, khususnya Twitter, masih menunjukkan adanya keraguan maupun penolakan yang dipengaruhi oleh faktor keamanan, efektivitas, kepercayaan terhadap pemerintah, hingga aspek keagamaan. Kondisi ini menimbulkan kebutuhan untuk melakukan analisis yang lebih komprehensif guna mengidentifikasi sentimen publik secara sistematis. Dalam ranah ilmu komputer, analisis sentimen berbasis machine learning dengan algoritma Multinomial Naive Bayes dipandang relevan karena mampu mengklasifikasikan komentar masyarakat menjadi sentimen positif atau negatif. Adapun tujuan penelitian ini adalah melakukan analisis sentimen masyarakat Indonesia terhadap vaksin AstraZeneca berdasarkan opini yang diunggah di media sosial Twitter, menerapkan dan melakukan evaluasi kinerja algoritma Multinomial Naive Bayes dalam melakukan klasifikasi sentimen publik terkait vaksin AstraZeneca. Dengan demikian, hasil penelitian ini diharapkan dapat memberikan kontribusi nyata dalam penyusunan strategi komunikasi pemerintah dan pemangku kebijakan, khususnya dalam lingkup penyebaran informasi di ruang digital.

II. TINJAUAN PUSTAKA

A. Analisis Sentimen

Analisis sentimen adalah proses mendeteksi, memahami, dan menginterpretasikan sentimen atau opini yang terkandung dalam teks, seperti artikel berita, ulasan produk, atau media sosial [8]. Teknologi analisis sentimen memungkinkan untuk secara otomatis mengidentifikasi apakah suatu teks mengandung sentimen positif, negatif, atau netral [9]. Analisis sentimen dilakukan dengan menggunakan berbagai teknik pemrosesan bahasa alami dan pembelajaran mesin, seperti klasifikasi teks dan analisis emosi, untuk mengekstrak dan menganalisis sentimen yang tersembunyi dalam teks [10]. Analisis sentimen dapat membantu perusahaan atau organisasi dalam memahami umpan balik pelanggan, memantau citra merek, dan mengidentifikasi tren pasar, sehingga memungkinkan mereka untuk mengambil tindakan yang sesuai dalam merespons perasaan dan opini yang tersebar di masyarakat [11]. Meskipun teknologi analisis sentimen menawarkan berbagai manfaat, termasuk pemahaman yang lebih baik tentang pandangan masyarakat dan tren pasar, penting untuk diingat bahwa hasilnya tidak selalu sempurna. Analisis sentimen sering kali rumit karena teks dapat berisi banyak nuansa dan konteks yang sulit diinterpretasikan oleh mesin secara akurat [12]. Selain itu, perbedaan budaya, bahasa, dan gaya penulisan dapat mempengaruhi hasil analisis sentimen. Oleh karena itu, penting untuk menggunakan teknologi analisis sentimen sebagai alat bantu, bukan pengganti, untuk penilaian manusia yang lebih mendalam dan kontekstual [13]. Dengan pendekatan yang hati-hati dan penggunaan yang bijak, analisis sentimen dapat menjadi sumber daya berharga dalam pengambilan keputusan dan strategi bisnis.

Berbagai penelitian terkait analisis sentimen COVID-19 telah banyak dilakukan dan tentu saja memiliki bermacam – macam perbedaan fokus, objek, metode, dan dataset, dan menjadi acuan dari penelitian kali ini. Penelitian pertama menyoroti masalah akurasi klasifikasi sentimen pada Twitter, dengan objek berupa cuitan COVID-19 yang dianalisis menggunakan lima algoritma machine learning (Logistic Regression, Random Forest, Multinomial Naive Bayes, SVM, dan Decision Tree) melalui dataset tweet yang dikumpulkan dengan Twint [14]. Penelitian kedua berfokus pada isu keterbatasan pemahaman topik dan sentimen dalam pemberitaan pandemi, dengan objek berupa lebih dari 100.000 berita dan judul artikel terkait COVID-19 dari berbagai negara, dianalisis menggunakan Top2Vec untuk pemodelan topik dan RoBERTa untuk klasifikasi sentimen [15]. Penelitian ketiga menekankan perlunya model analisis sentimen yang lebih ringan namun tetap berkualitas, dengan objek berupa tweet COVID-19 yang diproses menggunakan kombinasi BERT dan

deep CNN pada dataset skala besar [16]. Sementara itu, penelitian keempat mengangkat permasalahan beragamnya emosi publik tentang pandemi dan vaksin di Twitter, dengan objek berupa cuitan terkait COVID-19 yang diklasifikasikan ke dalam sentimen positif, negatif, dan netral menggunakan pendekatan machine learning dan deep learning, diuji pada dataset COVISenti, COVIDSenti_A, COVIDSenti_B, dan COVIDSenti_C dengan hasil akurasi tinggi hingga 96,66% [17].

B. *Multinomial Naive Bayes*

Multinomial Naïve Bayes merupakan varian dari model Naïve Bayes yang sangat berguna untuk klasifikasi teks [18]. Model ini menggunakan pendekatan supervised learning, yang berarti bahwa semua data pelatihan harus diberi label sebelum dijadikan acuan model. Berikut ini adalah rumus dasar tentang Multinomial Naïve Bayes.

$$P(w_i|c_j) = \frac{count(w_i,c_j)+1}{(\sum_{w\epsilon V} count(w,c_j))+|V|} \quad (1)$$

- $P(w_i|c_j)$: probabilitas dari sebuah kata i termasuk dalam kategori j.
- $count(w_i, c_j)$: Jumlah kata kueri yang muncul dalam suatu kelas atau kategori.
- Koefisien 1 ditambahkan untuk menghindari nilai nol.
- $\sum_{w\epsilon V} count(w, c_j)$: Jumlah semua kata dalam kelas atau kategori cj.
- $|V|$: Jumlah semua kata unik di semua kategori.

C. *Confussion Matrix*

Bagian ini menjelaskan perhitungan menggunakan Confusion Matrix, sebuah metode yang digunakan untuk mengevaluasi kinerja klasifikasi dengan membedakan antara prediksi yang benar dan salah. Dalam Confusion Matrix, empat metrik utama dihitung: Akurasi, Presisi, Recall, dan Spesifisitas, serta skor F1 [19]. Berikut ini adalah rumus untuk Akurasi, Presisi, Recall, dan Spesifisitas, beserta prioritasnya dalam Machine Learning.

1. Akurasi
   Pada akurasi membandingkan rasio True Positive (TP) dan True Negative (TN) pada seluruh data dengan rumus:
   $$Accuracy = \sum_{i=1}^{1}(\frac{\frac{TPi+TNi}{TPi+TNi+FPi+FNi}x100\%}{1}) \quad (2)$$
2. Presisi
   Pada presisi membandingkan rasio True Positive (TP) pada data yaitu prediksi positif dengan rumus:
   $$Precision = \frac{\sum_{i=1}^{1}(TPi)}{\sum_{i=1}^{1}(TPi+FPi)}x100\% \quad (3)$$
3. Recall
   Pada recall membandingkan rasio True Positive (TP) pada seluruh data yaitu True Positive dengan rumus:
   $$Recall = \frac{\sum_{i=1}^{1}(TPi)}{\sum_{i=1}^{1}(TPi+FNi)}x100\% \quad (4)$$
4. Skor F-1
   Skor F-1 merupakan hasil nilai rata-rata yang diperoleh dari presisi dan recall dengan menggunakan rumus:
   $$F1 = \frac{2\ x\ Precision\ x\ Recall}{Precision+Recall} \quad (5)$$

III. METODOLOGI PENELITIAN

Metodologi penelitian terdiri dari beberapa tahapan yaitu Crawling Data, Preprocessing, Feature Extraction, Classification.

Gambar 1 Metodologi Penelitian

A. *Crawling Data*

Tahap ini mempunyai tujuan yaitu mengumpulkan data terkait sentimen publik mengenai vaksin AstraZeneca di Indonesia dari media sosial Twitter. Alat dan teknik yang digunakan adalah Twitter API atau pustaka seperti Tweepy (untuk Python) sehingga dapat melakukan ekstrak tweet berdasarkan kata kunci seperti "AstraZeneca", "vaksin AstraZeneca", atau "vaksin COVID-19".

Langkah – langkah yang dilakukan adalah sebagai berikut:

- Setup API: Daftar dan autentikasi menggunakan API Twitter.
- Kumpulkan Data: Tentukan parameter pencarian seperti rentang waktu, bahasa (misalnya, Bahasa Indonesia), dan lokasi (Indonesia) untuk memastikan relevansi data.
- Simpan Data: Simpan hasil crawling ke dalam format CSV atau database untuk kemudahan akses dan analisis tahap selanjutnya.

B. *Preprocessing*

Tahap preprocessing bertujuan untuk membersihkan data dari noise dan menyiapkannya agar siap untuk proses analisis sentimen.

Langkah – langkah yang dilakukan adalah:

- Cleaning Data: Hapus URL, mentions, hashtags, dan karakter khusus yang tidak relevan.
- Case Folding: Ubah semua teks menjadi huruf kecil agar seragam.
- Tokenization: Pisahkan kata-kata dalam kalimat agar lebih mudah dianalisis.
- Stopword Removal: Hapus kata-kata umum yang tidak memiliki banyak makna seperti "dan", "atau", "yang" menggunakan daftar stopword bahasa Indonesia.
- Stemming: Lakukan stemming (misalnya, menggunakan Sastrawi untuk Bahasa Indonesia) untuk mengubah kata menjadi bentuk dasarnya.

C. *Feature Extraction*

Tujuan dari tahap feature extraction adalah mengubah data teks menjadi fitur numerik yang dapat digunakan oleh model

pembelajaran mesin. Teknik yang digunakan adalah metode TF-IDF (Term Frequency-Inverse Document Frequency) atau Bag of Words [20].

Langkah – langkah feature extraction adalah sebagai berikut:

- TF-IDF Calculation: Hitung bobot kata menggunakan metode TF-IDF untuk mengukur pentingnya kata dalam dataset.
- Vectorization: Ubah teks menjadi vektor numerik berdasarkan hasil TF-IDF atau Bag of Words.
- Data Splitting: Bagi data menjadi training set dan testing set.

D. Multinomial Naive Bayes Classifier

Classifier yang digunakan adalah multinomial naïve bayes, sehingga dari dataset yang digunakan diklasifikasikan nilai sentimen yang merepresentasikan opnini publik terhadap vaksin AstraZeneca (positif, negatif, atau netral).

Langkah – langkah yang dilakukan adalah:

- Training Model: Proses melatih model Multinomial Naive Bayes menggunakan training set yang telah diproses.
- Testing Model: Melakukan uji model menggunakan testing set untuk mengevaluasi performa klasifikasi.
- Evaluasi Model: Menggunakan metrik akurasi, presisi, recall, dan F1-score untuk menilai kinerja model dalam mengklasifikasikan sentimen [21].

IV. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Pada tahap pengumpulan data, penelitian ini melakukan pengambilan data berupa tweet dari platform X. Kata kunci yang digunakan untuk pencarian adalah "AstraZeneca". Dari proses pengumpulan data tersebut, diperoleh sejumlah 460 tweet. Data tweet yang terkumpul akan menjadi bahan analisis lebih lanjut untuk menjawab tujuan penelitian.

```
   ,created_at,favorite_count,full_text,id_str,image_url,in_reply_to_scr
1.79052E+18,Tue May 14 23:25:34 +0000 2024,20,Alasan Sosiolog Unair Sebut
1.79051E+18,Tue May 14 22:42:56 +0000 2024,16,. . menderita luka serius s
1.79051E+18,Tue May 14 22:42:55 +0000 2024,94,Raksasa farmasi Inggris-Swe
1.79043E+18,Tue May 14 17:18:35 +0000 2024,0,Sosiolog Unair Sebut Penarik
1.79042E+18,Tue May 14 16:47:44 +0000 2024,0,Semoga kalian yang menganjur
1.79042E+18,Tue May 14 16:18:38 +0000 2024,273,Ini old news dari Januari
1.79039E+18,Tue May 14 14:39:48 +0000 2024,1,Sehat sehat dan panjang umur
1.79039E+18,Tue May 14 14:37:14 +0000 2024,2,Perusahaan farmasi AstraZene
1.79037E+18,Tue May 14 13:33:26 +0000 2024,44,Gimana nolong rakyat yg sud
1.78967E+18,Tue May 14 12:31:41 +0000 2024,0,@MRpluviophile Waduh. Lupa m
1.79035E+18,Tue May 14 11:42:10 +0000 2024,1,Astrazeneca dan efek samping
1.79034E+18,Tue May 14 11:00:15 +0000 2024,0,apakah ak udah terkena efek
1.79034E+18,Tue May 14 10:59:23 +0000 2024,0,kepala rasa mcm kena cucuk c
1.79028E+18,Tue May 14 06:57:26 +0000 2024,1,Lancarnya peredaran darah me
1.79027E+18,Tue May 14 06:48:07 +0000 2024,0,Pfizer dan AstraZeneca Inves
1.79026E+18,Tue May 14 06:07:29 +0000 2024,0,Malina berkata situasi semas
1.78994E+18,Tue May 14 05:22:06 +0000 2024,0,@easy305 @PempheroMphande za
1.79025E+18,Tue May 14 05:02:04 +0000 2024,9,Kesan sampingan AZ cuma seme
1.79022E+18,Tue May 14 04:09:45 +0000 2024,14,@StopPlandemit Semoga dia u
1.47905E+18,Tue May 14 04:07:16 +0000 2024,6,@tirta_cipeng Semoga Lo udah
1.79017E+18,Mon May 13 23:44:58 +0000 2024,1,Penerima vaksin AstraZeneca
```

Gambar 2 Data Hasil Scrapping Twitter

Setelah melakukan scraping data tweet, langkah selanjutnya adalah pelabelan untuk mengukur seberapa akurat metode yang digunakan. Pelabelan dilakukan dengan membagi data menjadi tiga kelas: negatif, netral, dan positif. Proses ini dilakukan secara manual pada 460 tweet. Setelah pelabelan manual selesai, validasi dilakukan oleh ahli bahasa untuk memastikan keakuratannya. Setelah selesai melakukan pelabelan sentimen pada data Tweet, ditemukan bahwa terdapat 142 Tweet dengan sentimen Negatif, 162 Tweet dengan sentimen Netral, dan 156 Tweet dengan sentimen Positif. Hanya tweet dengan sentimen Negatif dan Positif yang akan digunakan untuk implementasi metode Naive Bayes.

```
1  No,Tweets,Label
2  1,Alasan Sosiolog Unair Sebut Penarikan Vaksin AstraZeneca Bisa
3  2,. . menderita luka serius setelah menerima vaksin COVID ini. A
4  3,Raksasa farmasi Inggris-Swedia AstraZeneca telah mulai mencabu
5  4,Sosiolog Unair Sebut Penarikan Vaksin AstraZeneca Perburuk Sit
6  5,Semoga kalian yang menganjurkan BISA BERTANGGUNG JAWAB atas KO
7  6,Ini old news dari Januari 2024 tapi waktu itu tidak banyak yan
8  7,Sehat sehat dan panjang umur untuk aku dan kalian semua yang p
9  8,Perusahaan farmasi AstraZeneca mengakui vaksin yang dikembangk
10 9,Gimana nolong rakyat yg sudah kena KIPI ASTRAZENECA?,Netral
```

Gambar 3 Hasil Pelabelan

B. Preprocessing

Proses Preprocessing bertujuan untuk mengubah teks mentah menjadi bentuk yang lebih mudah diproses oleh metode Naive Bayes. Tahapan Preprocessing yang dilakukan meliputi Case Folding, Data Cleaning, Normalisasi, Stopwords Removal, Stemming, dan Tokenisasi.

Case folding adalah proses mengubah semua huruf dalam teks menjadi huruf kecil untuk memastikan keseragaman. Hasil case folding ditunjukkan pada tabel 1.

Tabel 1 Hasil Case Folding

Tweets	Case folding
Vaksin AstraZeneca sangat aman dn efek sampingnya ringan.	vaksin astrazeneca sangat aman dn efek sampingnya ringan.
Efek samping vaksin AstraZeneca membuat sy demam tinggi.	efek samping vaksin astrazeneca membuat sy demam tinggi.
Aman dn bisa ditoleransi, vaksin AstraZeneca jd pilihan yang tepat	aman dn bisa ditoleransi, vaksin astrazeneca jd pilihan yang tepat
Aman sih katanya, tp efek samping vaksin AstraZeneca bs bikin khawatir.	aman sih katanya, tp efek samping vaksin astrazeneca bs bikin khawatir.

Tabel 2 menunjukkan hasil dari data cleaning, Data Cleaning melibatkan penghapusan karakter-karakter yang tidak relevan seperti angka, tanda baca, dan simbol.

Tabel 2 Hasil Data Cleaning

Case folding	Data Cleaning
vaksin astrazeneca sangat aman dn efek sampingnya ringan.	vaksin astrazeneca sangat aman dn efek sampingnya ringan
efek samping vaksin astrazeneca membuat sy demam tinggi.	efek samping vaksin astrazeneca membuat sy demam tinggi
aman dn bisa ditoleransi, vaksin astrazeneca jd pilihan yang tepat	aman dn bisa ditoleransi vaksin astrazeneca jd pilihan yang tepat
aman sih katanya, tp efek samping vaksin astrazeneca bs bikin khawatir.	aman sih katanya tp efek samping vaksin astrazeneca bs bikin khawatir

Normalisasi adalah proses mengganti kata-kata informal atau bahasa gaul dengan bentuk yang lebih baku, ditunjukkan

pada tabel 3.

Stopwords Removal adalah kata-kata umum yang biasanya tidak membawa banyak informasi penting, seperti "dan", "yang", "di",”tidak”,”bisa”,sebut”,”ini”,tapi”, dsb. Penghapusan stopwords membantu mengurangi dimensi data. Hasil Stopwords removal ditunjukkan pada tabel 4.

Stemming adalah proses mengubah kata-kata ke bentuk dasar atau akarnya. Misalnya, kata-kata seperti "berlari", "berlarian", dan "pelari" diubah menjadi kata dasar "lari". Hasil proses stemming disajikan pada tabel 5.

Tabel 6 menunjukkan hasil tokenisasi. Tokenisasi adalah proses memecah teks menjadi unit-unit kata yang lebih kecil, disebut token. Setiap kata dalam teks dianggap sebagai satu token.

Tabel 3 Hasil Normalisasi

Data Cleaning	Normalisasi
vaksin astrazeneca sangat aman dn efek sampingnya ringan	vaksin astrazeneca sangat aman dan efek sampingnya ringan
efek samping vaksin astrazeneca membuat sy demam tinggi	efek samping vaksin astrazeneca membuat saya demam tinggi
aman dn bisa ditoleransi vaksin astrazeneca jd pilihan yang tepat	aman dan bisa ditoleransi vaksin astrazeneca jadi pilihan yang tepat
aman sih katanya tp efek samping vaksin astrazeneca bs bikin khawatir	aman sih katanya tapi efek samping vaksin astrazeneca bisa bikin khawatir

Tabel 4 Hasil Stopwords Removal

Normalisasi	Stopword Removal
vaksin astrazeneca sangat aman dan efek sampingnya ringan	vaksin astrazeneca sangat aman efek sampingnya ringan
efek samping vaksin astrazeneca membuat saya demam tinggi	efek samping vaksin astrazeneca membuat demam tinggi
aman dan bisa ditoleransi vaksin astrazeneca jadi pilihan yang tepat	aman ditoleransi vaksin astrazeneca jadi pilihan tepat
aman sih katanya tapi efek samping vaksin astrazeneca bisa bikin khawatir	aman katanya tapi efek samping vaksin astrazeneca bisa bikin khawatir

Tabel 5 Hasil Stemming

Stopword Removal	Stemming
vaksin astrazeneca sangat aman efek sampingnya ringan	vaksin astrazeneca sangat aman efek samping ringan
efek samping vaksin astrazeneca membuat demam tinggi	efek samping vaksin astrazeneca buat demam tinggi
aman ditoleransi vaksin astrazeneca jadi pilihan tepat	aman toleransi vaksin astrazeneca jadi pilih tepat
aman katanya tapi efek samping vaksin astrazeneca bisa bikin khawatir	aman kata tapi efek samping vaksin astrazeneca bisa bikin khawatir

Tabel 6 Hasil Tokenisasi

Stemming	Tokenisasi
vaksin *astrazeneca* sangat aman efek samping ringan	["vaksin", "*astrazeneca*", "sangat", "aman", "efek", "samping", "ringan"]
efek samping vaksin *astrazeneca* buat demam tinggi	["efek", "samping", "vaksin", "*astrazeneca*", "buat", "demam", "tinggi"]
aman toleransi vaksin *astrazeneca* jadi pilih tepat	["aman","toleransi", "vaksin", "*astrazeneca*", "jadi", "pilih", "tepat"]
aman kata tapi efek samping vaksin *astrazeneca* bisa bikin khawatir	["aman", "kata", "tapi", "efek", "samping", "vaksin", "*astrazeneca*", "bisa", "bikin", "khawatir"]

C. Hasil Implementasi metode Naive Bayes

Metode Naive Bayes diimplementasikan dengan menggunakan bahasa pemrograman python.

```
# Inisialisasi dan pelatihan model dengan laplace smoothing
class NaiveBayesWithLaplace:
    def fit(self, X, y):
        self.classes = np.unique(y)
        self.feature_count = X.shape[1]
        self.class_count = len(self.classes)

        # Initialize count matrices
        self.class_prior = np.zeros(self.class_count)
        self.likelihood = np.zeros((self.class_count, self.feature_count))

        for i, c in enumerate(self.classes):
            X_c = X[y == c]
            self.class_prior[i] = (X_c.shape[0] + 1) / (X.shape[0] + self.class_count)
            self.likelihood[i, :] = (X_c.sum(axis=0) + 1) / (X_c.sum() + self.feature_count)

    def predict(self, X):
        log_prob = np.log(self.class_prior) + X @ np.log(self.likelihood.T)
        return self.classes[np.argmax(log_prob, axis=1)]

nb_model_laplace = NaiveBayesWithLaplace()
nb_model_laplace.fit(X_train_smote, y_train_smote)
```

Gambar 4 Implementasi Naïve Bayes dengan Python

Pada proses implementasi metode Naive Bayes, dilakukan Oversampling untuk mengatasi ketidakseimbangan kelas dalam data latih. Oversampling dilakukan menggunakan teknik SMOTE (Synthetic Minority Over-sampling Technique) dari library imbalanced-learn untuk memastikan bahwa setiap kelas memiliki representasi yang seimbang dalam data latih.

```
# Oversampling pada data pelatihan menggunakan SMOTE
smote = SMOTE(random_state=42)
X_train_resampled, y_train_resampled = smote.fit_resample(X_train_vectorized, y_train)
```

Gambar 5 Proses SMOTE

Setelah proses Oversampling, teks diubah menjadi fitur-fitur numerik menggunakan TF-IDF (Term Frequency-Inverse Document Frequency) Vectorizer.

```
# Ekstraksi fitur dengan TF-IDF
tfidf_vectorizer = TfidfVectorizer(tokenizer=lambda text: text.split())

X_train_vectorized = tfidf_vectorizer.fit_transform(X_train)
X_val_vectorized = tfidf_vectorizer.transform(X_val)
X_test_vectorized = tfidf_vectorizer.transform(X_test)
```

Gambar 6 TF-IDF

Tabel 7 Hasil Uji

Pengukuran	*Jumlah*
True Positives (TP)	133
True Negatives (TN)	126
False Positives (FP)	16
False Negatives (FN)	23

Model Naïve Bayes yang dilatih menunjukkan kinerja yang cukup baik berdasarkan evaluasi terhadap keseluruhan data. Berdasarkan Confusion Matrix, terdapat 126 prediksi benar untuk kelas negatif dan 133 prediksi benar untuk kelas positif, dengan 16 prediksi salah untuk kelas negatif dan 23 prediksi

salah untuk kelas positif.

V. KESIMPULAN DAN SARAN

Performa model Naive Bayes dalam mengklasifikasikan sentimen masyarakat terhadap efek samping vaksin AstraZeneca di X cukup baik berdasarkan evaluasi yang dilakukan.

Model ini mencapai akurasi sekitar 86.91%, yang menunjukkan bahwa hampir 87% dari semua prediksi yang dibuat oleh model adalah benar. Dalam hal presisi, model memiliki nilai sekitar 89.26%, yang berarti bahwa sekitar 89% dari prediksi positif adalah benar. Recall model adalah sekitar 85.26%, yang menunjukkan bahwa model mampu mengidentifikasi lebih dari 85% dari total data positif.

Interpretasi sentimen dari hasil uji menunjukkan bahwa sentimen publik hampir seimbang, menunjukkan adanya dukungan dan kekhawatiran yang relatif seimbang terhadap vaksin COVID-19 AstraZeneca.

REFERENSI

[1] M. A. Shereen, S. Khan, A. Kazmi, N. Bashir, and R. Siddique, "COVID-19 infection: Emergence, transmission, and characteristics of human coronaviruses," *J. Adv. Res.*, vol. 24, pp. 91–98, 2020, doi: https://doi.org/10.1016/j.jare.2020.03.005.

[2] A. Sanyaolu *et al.*, "Global Pandemicity of COVID-19: Situation Report as of June 9, 2020," *Infect. Dis. Res. Treat.*, vol. 14, p. 1178633721991260, 2021, doi: 10.1177/1178633721991260.

[3] I. Hikmawati and R. Setiyabudi, "Epidemiology of COVID-19 in Indonesia: common source and propagated source as a cause for outbreaks," *J. Infect. Dev. Ctries.*, vol. 15, no. 05 SE-Coronavirus Pandemic, pp. 646–652, May 2021, doi: 10.3855/jidc.14240.

[4] I. K. TUNAS, D. A. Y. U. A. SRI LAKSEMI, I. P. E. K. A. WIDYADHARMA, and L. U. H. P. R. SUNDARI, "THE EFFICACY OF COVID-19 VACCINE AND THE CHALLENGE IN IMPLEMENTING MASS VACCINATION IN INDONESIA," *Int. J. Appl. Pharm.*, vol. 13, no. 4 SE-Review Article(s), pp. 74–76, Jul. 2021, doi: 10.22159/ijap.2021v13i4.41270.

[5] Q. Huang and J. Yan, "SARS-CoV-2 virus: Vaccines in development.," *Fundam. Res.*, vol. 1, no. 2, pp. 131–138, Mar. 2021, doi: 10.1016/j.fmre.2021.01.009.

[6] B. C. Fialho, L. Gauss, P. F. Soares, M. Z. Medeiros, and D. P. Lacerda, "Vaccine Innovation Meta-Model for Pandemic Contexts," *J. Pharm. Innov.*, vol. 18, no. 3, pp. 1145–1193, 2023, doi: 10.1007/s12247-023-09708-7.

[7] R. Septinia and D. Hasmono, "Covid-19 and Its Vaccine Development: A Narrative Review," *J. Sains dan Kesehat.*, vol. 3, no. 6, pp. 917–929, 2021, [Online]. Available: https://jsk.ff.unmul.ac.id/index.php/JSK/article/view/520

[8] F. Aftab *et al.*, "A Comprehensive Survey on Sentiment Analysis Techniques," *Int. J. Technol.*, vol. 14, no. 6, pp. 291–319, 2023, doi: https://doi.org/10.14716/ijtech.v14i6.6632.

[9] Tanvi, K. B. Somani, and P. Bhargav, "A Novel Approach of Deep Learning Using Sentiment Analysis in Python," *Tuijin Jishu/Journal Propuls. Technol.*, vol. 44, no. 4, 2023, doi: https://doi.org/10.52783/tjjpt.v44.i4.1617.

[10] W. Etaiwi, D. Suleiman, and A. Awajan, "Deep Learning Based Techniques for Sentiment Analysis: A Survey," *Informatica*, vol. 45, no. 7, 2021, doi: https://doi.org/10.31449/inf.v45i7.3674.

[11] N. Nasrabadi, H. Wicaksono, and O. Fatahi Valilai, "Shopping marketplace analysis based on customer insights using social media analytics," *MethodsX*, vol. 13, p. 102868, 2024, doi: https://doi.org/10.1016/j.mex.2024.102868.

[12] B. Wadhvani *et al.*, "Leading-edge Sentiment Analysis: A Survey of Application Context, Challenges and Advanced Techniques," *Recent Advances in Computer Science and Communications*, vol. 18, no. 6. pp. 56–71, 2025. doi: http://dx.doi.org/10.2174/0126662558312258240911111028.

[13] Abhik Roy and Karen E Rambo-Hernandez, "There's So Much to Do and Not Enough Time to Do It! A Case for Sentiment Analysis to Derive Meaning From Open Text Using Student Reflections of Engineering Activities," *Am. J. Eval.*, vol. 42, no. 4, pp. 559–576, Oct. 2021, doi: 10.1177/1098214020962576.

[14] D. Dangi, D. K. Dixit, and A. Bhagat, "Sentiment analysis of COVID-19 social media data through machine learning," *Multimed. Tools Appl.*, vol. 81, no. 29, pp. 42261–42283, 2022, doi: 10.1007/s11042-022-13492-w.

[15] P. Ghasiya and K. Okamura, "Investigating COVID-19 News Across Four Nations: A Topic Modeling and Sentiment Analysis Approach," *IEEE Access*, vol. 9, pp. 36645–36656, 2021, doi: 10.1109/ACCESS.2021.3062875.

[16] J. H. Joloudari *et al.*, "BERT-deep CNN: state of the art for sentiment analysis of COVID-19 tweets," *Soc. Netw. Anal. Min.*, vol. 13, no. 1, p. 99, 2023, doi: 10.1007/s13278-023-01102-y.

[17] Z. Jalil *et al.*, "COVID-19 Related Sentiment Analysis Using State-of-the-Art Machine Learning and Deep Learning Techniques," *Front. Public Heal.*, vol. Volume 9-2021, 2022, [Online]. Available: https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2021.812735

[18] N. Hayatin, S. Alias, L. P. Hung, and M. S. Sainin, "SENTIMENT ANALYSIS BASED ON PROBABILISTIC CLASSIFIER TECHNIQUES IN VARIOUS INDONESIAN REVIEW DATA," *Jordanian J. Comput. Inf. Technol.*, vol. 08, no. 03, pp.

271–281, 2022, doi: 10.5455/jjcit.71-1646912715.
[19] D. Chicco and G. Jurman, "An Invitation to Greater Use of Matthews Correlation Coefficient in Robotics and Artificial Intelligence," *Front. Robot. AI*, vol. Volume 9-, 2022, doi: 10.3389/frobt.2022.876814.
[20] J. Ni, Y. Cai, G. Tang, and Y. Xie, "Collaborative Filtering Recommendation Algorithm Based on TF-IDF and User Characteristics," *Applied Sciences*, vol. 11, no. 20. 2021. doi: 10.3390/app11209554.
[21] S. Silva, A. Seara Vieira, P. Celard, E. L. Iglesias, and L. Borrajo, "A Query Expansion Method Using Multinomial Naive Bayes," *Applied Sciences*, vol. 11, no. 21. 2021. doi: 10.3390/app112110284.