Analisis Sentimen Aplikasi Gojek Pada Twitter Menggunakan Algoritma Naive Bayes

Adhi Dwi Saputra[1] , Daniel Prasetio Budiman[2], Revano Maliq Reynanda[3], Anggraini Puspita Sari[4]
[1,2,3,4]Universitas Pembangunan Nasional "Veteran" Jawa timur, Kota Surabaya
Email:
22081010287@student.upnjatim.ac.id[1], 22081010304@student.upnjatim.ac.id[2],
22081010334@student.upnjatim.ac.id[3], anggraini.puspita.if@upnjatim.ac.id[4]
Email : corresponding 22081010304@student.upnjatim.ac.id

Abstrak

Media sosial telah menjadi platform penting bagi masyarakat untuk menyampaikan opini tentang produk dan layanan. Salah satu aplikasi yang populer di Indonesia adalah Gojek, yang banyak dibicarakan di Twitter. Penelitian ini bertujuan untuk menganalisis sentimen pengguna Twitter terhadap layanan Gojek menggunakan algoritma Naive Bayes. Tahapan penelitian meliputi pengumpulan data tweet, pelabelan data, praproses data, transformasi dengan TF-IDF, klasifikasi menggunakan Naive Bayes, evaluasi model, dan visualisasi hasil. Dalam penelitian ini, dibandingkan beberapa perbandingan pembagian data, yaitu 80:20, 75:25, 70:30, dan hasil menunjukkan bahwa persentase yang paling besar adalah 80:20, yang kemudian digunakan dalam penelitian ini. Hasil menunjukkan bahwa model Naive Bayes mampu mengklasifikasikan sentimen dengan akurasi 84,21%, presisi 86,67%, dan recall 92,86%. Penelitian ini bertujuan memberikan kontribusi akademis dalam bidang analisis sentimen dan pengolahan bahasa alami, serta memperkaya literatur terkait dengan studi kasus aplikasi layanan digital di Indonesia. Selain itu, penelitian ini membuka peluang untuk pengembangan lebih lanjut dengan algoritma lain dan analisis yang lebih mendalam.

Kata kunci: Analisis Sentimen, Naive Bayes, Twitter, Gojek, TF-IDF

Abstract

Social media has become a vital platform for individuals to express their opinions about products and services. One of the most prominent applications in Indonesia is Gojek, which is frequently discussed on Twitter. This research aims to analyze Twitter users' sentiments towards Gojek's services utilizing the Naive Bayes algorithm. The study compared various data split ratios, specifically 80:20, 75:25 and 70:30, determining that the 80:20 split yielded the most favorable results, thus it was employed in this research. The findings indicate that the Naive Bayes model achieved a sentiment classification accuracy of 84,21%, a precision of 86,67%, and a recall of 92,86%. This study aims to contribute academically to the fields of sentiment analysis and natural language processing, while also enriching the existing literature related to digital service applications in Indonesia. Furthermore, this research presents opportunities for future development using alternative algorithms and more comprehensive analysis.

Keywords: Sentiment Analysis, Naive Bayes, Twitter, Gojek, TF-IDF

PENDAHULUAN

Media sosial telah menjadi platform penting bagi masyarakat untuk berkomunikasi dan bertukar informasi di era saat ini. Salah satu platform media sosial yang populer di Indonesia adalah Twitter, yang memungkinkan pengguna untuk membagikan pesan singkat yang disebut tweet. Tweet-tweet ini dapat berisi berbagai macam informasi, termasuk opini dan ulasan tentang produk dan layanan (Pak & Paroubek, 2010). Analisis sentimen terhadap tweet-tweet ini dapat memberikan wawasan berharga tentang persepsi dan pengalaman pengguna terhadap produk atau layanan tertentu. Menurut Al Haromainy, M. M., Prasetya, D. A., & Sari, A. P. (2023), analisis sentimen merupakan salah satu aplikasi penting dalam pengolahan bahasa alami (NLP) yang bertujuan untuk mengekstrak opini atau sentimen dari teks. Gojek merupakan aplikasi transportasi online yang telah menjadi bagian dari kehidupan masyarakat Indonesia. Gojek menawarkan berbagai layanan, seperti transportasi roda dua, roda empat, pesan-antar makanan, dan lain sebagainya. Popularitas Gojek mendorong banyak pengguna Twitter untuk membagikan pengalaman mereka menggunakan aplikasi tersebut. Dengan menganalisis sentimen dari tweet-tweet ini, perusahaan dapat memahami lebih baik kebutuhan dan kepuasan pelanggan, serta mengidentifikasi area yang perlu diperbaiki (Medhat et al., 2014).

Analisis sentimen pada tweet-tweet terkait Gojek dapat memberikan gambaran tentang bagaimana pengguna merasa tentang berbagai aspek layanan yang ditawarkan oleh Gojek, seperti keandalan, kecepatan, dan kualitas layanan. Informasi ini sangat berharga bagi Gojek untuk meningkatkan layanannya dan tetap kompetitif di pasar yang semakin kompetitif. Metode Naive Bayes memiliki beberapa kelebihan yang menjadikannya populer dalam analisis sentimen. Algoritma ini sederhana dan mudah diimplementasikan, serta mampu bekerja dengan baik dalam mengolah data teks yang bersifat tidak terstruktur (Kowsari et al., 2019). Menurut Rani dan Kumar (2017), Algoritma Naive Bayes menghitung probabilitas suatu teks termasuk dalam kelas tertentu dengan menganalisis frekuensi kata-kata dalam teks tersebut. Model ini menghitung probabilitas suatu teks termasuk dalam kategori tertentu (misalnya positif, negatif, atau netral) berdasarkan kemunculan kata-kata dalam teks tersebut. Meskipun asumsi independensi kata dalam Naive Bayes jarang sepenuhnya akurat, algoritma ini tetap memberikan hasil yang memuaskan dalam banyak aplikasi praktis.

Penelitian ini akan menggunakan algoritma Naive Bayes untuk mengklasifikasikan sentimen tweet pengguna Twitter terhadap aplikasi Gojek. Langkah-langkah yang akan dilakukan meliputi pengumpulan data tweet yang terkait dengan Gojek, praproses data untuk menghapus noise dan fitur yang tidak relevan, transformasi teks menjadi representasi numerik, pelatihan model Naive Bayes, dan evaluasi kinerja model. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi dalam memahami opini publik terhadap layanan Gojek serta membantu perusahaan dalam pengambilan keputusan yang lebih baik. Melalui penelitian ini, Gojek dapat memperoleh wawasan yang lebih mendalam tentang sentimen pengguna terhadap layanannya. Misalnya, dengan mengetahui aspek-aspek tertentu yang sering mendapatkan sentimen negatif, Gojek dapat fokus pada peningkatan aspek-aspek tersebut. Sebaliknya, aspek yang mendapatkan sentimen positif dapat dipertahankan atau ditingkatkan lebih lanjut untuk memastikan kepuasan pelanggan yang tinggi.

Dengan demikian, analisis sentimen dapat menjadi alat yang sangat berguna bagi perusahaan dalam strategi peningkatan layanan dan kepuasan pelanggan. Secara keseluruhan, penelitian ini bertujuan untuk memberikan pemahaman yang lebih baik tentang bagaimana pengguna Twitter

merespons aplikasi Gojek. Dengan menggunakan algoritma Naive Bayes, penelitian ini tidak hanya berfokus pada klasifikasi sentimen, tetapi juga memberikan wawasan yang dapat diimplementasikan oleh Gojek untuk meningkatkan layanannya. Penelitian ini diharapkan dapat menjadi referensi bagi studi-studi selanjutnya dalam analisis sentimen menggunakan media sosial dan algoritma pembelajaran mesin. Dengan adanya pemahaman yang lebih baik ini, Gojek diharapkan dapat terus berinovasi dan memenuhi ekspektasi pengguna dalam menghadapi persaingan yang semakin ketat di industri layanan digital.

METODE

Penelitian ini menggunakan metode klasifikasi Naive Bayes sebagai landasan utamanya. Proses penelitian mencakup beberapa tahap, yaitu Literature Review, Crawling Data, Labelling Data, Preprocessing, Transformation, Naïve Bayes Classification, dan Evaluation. Penjelasan dari tahapan metodologi penelitian dapat dilihat pada Gambar 1.

Gambar 1: Metodologi Penelitian

Penjelasan Gambar 1 adalah sebagai berikut:

1) *Literature Review* (Kajian Literatur) merupakan langkah awal yang penting dalam penelitian yang bertujuan untuk memetakan pengetahuan yang telah ada terkait dengan topik penelitian (Sukardi, 2014). Kajian literatur bertujuan guna memahami lebih dalam mengenai dasar analisis sentimen, data mining dengan menggunakan jurnal, artikel, karya ilmiah sebagai referensi

2) *Crawling Data* (Pengambilan Data) tahap ini bertujuan untuk mengumpulkan data. Data diunduh dari platform Twitter menggunakan alat Tweet-Harvest, dengan fokus pada tweet yang berhubungan dengan Gojek.

3) *Labelling Data* (Pemberian Label pada Data) merupakan proses memberi label atau kategori pada data. Mengumpulkan data mengenai sentimen terhadap Gojek dengan memberi label data sebagai positif atau negatif.

4) *Preprocessing* (Pra-pemrosesan) adalah langkah awal dalam analisis data di mana data mentah dibersihkan dan dipersiapkan untuk analisis lebih lanjut. Langkah ini krusial untuk meningkatkan kualitas data yang berdampak signifikan pada kinerja dan akurasi model machine learning (Hodge & Austin, 2021). Terdapat beberapa tahap preprocessing yang dilakukan dalam penelitian ini yakni *case folding*, *cleansing*, *stemming*, dan *stopwords filtering*.

5) *Transformation* (Transformasi) yaitu pada tahap ini Transformasi data melibatkan mengubah struktur atau format data menjadi bentuk yang lebih sesuai untuk analisis atau pemodelan. Pada penelitian ini menggunakan metode TF-IDF. Metode TF-IDF (Term Frequency-Inverse Document Frequency) digunakan untuk mengekstraksi fitur dari teks dengan memberikan bobot pada kata-kata berdasarkan frekuensi kemunculannya dalam suatu dokumen relatif terhadap frekuensi kemunculannya dalam korpus. TF-IDF memainkan peran penting dalam analisis sentimen dengan membantu mengidentifikasi istilah kunci yang membedakan kelas sentimen. Pra-pemrosesan yang tepat, termasuk normalisasi teks dan ekstraksi fitur, meningkatkan kinerja TF-IDF dan berkontribusi pada akurasi yang lebih baik dalam mengklasifikasikan sentimen sebagai positif atau negatif (Taneja, 2024). Rumus TF-IDF terdiri dari dua komponen: Term Frequency (TF) dan Inverse Document Frequency (IDF). TF mengukur frekuensi kemunculan sebuah kata dalam sebuah dokumen. Semakin sering sebuah kata muncul dalam dokumen, semakin tinggi nilai TF-nya (Alfarizi et al., 2022). IDF

mengukur seberapa umum atau jarang sebuah kata muncul di seluruh kumpulan dokumen. Semakin jarang sebuah kata muncul dalam kumpulan dokumen, semakin tinggi nilai IDF-nya. Nilai TF-IDF dihitung dengan mengalikan nilai TF dengan nilai IDF untuk setiap kata dalam dokumen.

6) Naïve Bayes Classification (Klasifikasi Naïve Bayes) adalah tahap pengklasifikasian menggunakan metode statistik yang memprediksi peluang keanggotaan kelas, seperti probabilitas tupel tertentu untuk bergabung dengan kelas tertentu. Metode ini mengasumsikan bahwa semua atribut bersifat independen dan dalam penerapannya menawarkan tingkat akurasi dan kecepatan yang cukup tinggi. Pada tahap pengujian, data dibagi secara acak menjadi dua kategori, yaitu data latih dan data uji. Menurut buku Applied Predictive Modeling dalam pemodelan prediktif, pembagian data adalah langkah penting dalam menilai kinerja model. Data latih digunakan untuk membangun model yang dapat memprediksi apakah suatu tweet termasuk dalam kategori positif atau negatif. Sementara itu, data uji digunakan untuk mengukur akurasi model klasifikasi. Dalam penelitian ini, rasio antara data pelatihan dan data pengujian bervariasi, termasuk rasio 80:20, 75:25, dan 70:30.

7) *Evaluation* (Evaluasi) adalah proses krusial dalam menilai kinerja model atau hasil analisis data yang dilakukan untuk memastikan efektivitas dan akurasi dari sistem yang dikembangkan. Evaluasi melibatkan penerapan *confusion matrix* untuk mengukur berbagai metrik kinerja, termasuk akurasi, presisi, dan recall. Akurasi mengukur proporsi total prediksi yang benar dari keseluruhan prediksi yang dibuat, sedangkan presisi menilai proporsi prediksi positif yang benar dibandingkan dengan total prediksi positif yang dibuat. Recall, mengukur proporsi prediksi positif yang benar dari total kasus positif yang sebenarnya.

HASIL DAN PEMBAHASAN

Hasil

Pada tahap ini, hasil analisis akan dipaparkan menggunakan metodologi yang telah dijelaskan sebelumnya, yang meliputi Crawling Data, Labelling Data, Preprocessing Data, Transformation, Naïve Bayes Classification, dan Evaluation.

1. *Crawling Data*

Data diambil menggunakan API (Application Programming Interface) Twitter dengan bantuan alat Tweet-Harvest pada tanggal 6 Mei 2024. Pengumpulan data mencakup periode satu tahun ke belakang, mulai dari 6 Mei 2023. Proses pengambilan data diawali dengan menginput token autentikasi Twitter melalui akun yang sudah terautentikasi. Setelah itu, tools akan mencari kata kunci yang diinginkan serta menentukan jumlah tweet yang akan diambil. Dari hasil pengumpulan data, didapatkan 1000 tweet dengan kata kunci "Gojek". Proses pencarian dapat dilihat pada Gambar 2.

```
filename = 'Gojek.csv'
search_keyword = 'Gojek lang:id'
limit = 1000
since_date = '2023-06-06'
until_date = '2024-06-06'

!npx --yes tweet-harvest@2.6.1 -o "{filename}" -s "{search_keyword}" -l {limit} --token {twitter_auth_token}
```

```
Filling in keywords: Gojek lang:id

-- Scrolling... (1)

Your tweets saved to: /content/tweets-data/Gojek.csv
Total tweets saved: 19

-- Scrolling... (1)

Your tweets saved to: /content/tweets-data/Gojek.csv
Total tweets saved: 39

-- Scrolling... (1)
```

Gambar 2. Proses Pencarian Kata Kunci

Setelah proses pencarian kata kunci selesai dilakukan, hasil perolehan data dapat dilihat pada Gambar 3.

	conversation_id_str	created_at	favorite_count	full_text	id_str	image_url	in_reply_to_screen_name	lang
0	1550830954624606209	Sat Jul 23 13:11:27 +0000 2022	27192	w setiap diajak ngobrol abang gojek: https://t...	1550830954624606209	https://pbs.twimg.com/media/FYWoMVEAAMzce.jpg	NaN	in
1	1231238077875183616	Sat Feb 22 15:27:26 +0000 2020	35417	Kode promo Gojek Cobain aja https://t.co/0NAtd...	1231238077875183616	https://pbs.twimg.com/media/ERY9YoAU1AAPQzsW.jpg	NaN	in
2	1169893656700293122	Fri Sep 06 15:19:59 +0000 2019	20128	Fitur terbaru gojek https://t.co/NSxHNEgNiV	1169893656700293122	https://pbs.twimg.com/media/EDynBNTU0AAAqvO.jpg	NaN	in
3	1211588033398460419	Mon Dec 30 10:01:12 +0000 2019	3016	om ngakak bgt"d ini ma bkan mesan gojek tapi me...	1211588033398460419	https://pbs.twimg.com/media/ENBk6PjVAABou4.jpg	NaN	in
4	1260894587347963905	Thu May 14 11:27:50 +0000 2020	6007	jdi ak mau cerita tdi sore ak sma kakaku cod o...	1260894587347963905	https://pbs.twimg.com/media/EX-Y3OqWAAEvdaP.jpg	NaN	in

Gambar 3. Hasil Crawling Data

2. *Labelling Data*

Pada langkah ini, data yang telah diperoleh akan diberi label secara manual. Dari 1.000 data yang didapatkan, hanya 92 data saja yang benar-benar dapat dikategorikan sebagai sentimen setelah melalui proses filtering. Label yang digunakan adalah 1 untuk sentimen positif dan 0 untuk sentimen negatif. Untuk hasil pelabelan sentimen lebih akurat, penilaian label sentimen sangatlah penting dilakukan langsung oleh manusia. Hasil dari pelabelan sentimen manual terdokumentasikan dalam Tabel 1.

Tabel 1. Hasil Pelabelan Data

Text	Label
Sering naik Gojek, nyaman	1
woy gojek kerja yang bener	0

Setelah proses pemberian label pada data selesai, ditemukan sebanyak 92 data yang memiliki sentimen positif dan negatif. Informasi terperinci mengenai jumlah tweet yang tergolong positif dan negatif terdapat pada Tabel 2.

Tabel 2. Jumlah Tweet Positif dan Negatif

Tweet	Jumlah Data
Positif	56
Negatif	36

3. *Preprocessing Data*

Tahap selanjutnya adalah preprocessing, di mana data diproses secara sistematis agar siap untuk analisis lebih lanjut. Berikut adalah hasil dari tahapan preprocessing:

a. *Case Folding*

Semua teks atau kalimat dalam data tweet ditransformasi menjadi huruf kecil melalui proses case folding. Tujuannya adalah untuk mempermudah pencocokan dokumen dengan ukuran huruf yang seragam. Hasil dari proses ini dapat dilihat dalam Tabel 3.

Tabel 3. Hasil Proses Case Folding

Sebelum
DANGGG... DPT DRY TEXT DRI GOJEK....
Sesudah
danggg... dpt dry text dri gojek....

b. *Cleansing*

Proses pembersihan dilakukan untuk menghapus karakter-karakter yang tidak diperlukan seperti angka, tanda baca, URL, emotikon, username Twitter, dan elemen lainnya. Hasil dari tahap ini dapat dilihat dalam Tabel 4.

Tabel 4. Hasil Proses Cleansing

Sebelum
@sassygurlya dry text si mang gojek
Sesudah
dry text si mang gojek

c. *Stemming*

Stemming adalah tahap di mana kata-kata yang memiliki imbuhan diubah menjadi bentuk dasarnya. Dalam tahap ini, kata-kata yang berimbuhan akan disederhanakan menjadi bentuk dasarnya. Hasil dari proses stemming dapat ditemukan dalam Tabel 5.

Tabel 5. Hasil Proses Stemming

Sebelum
orang yang gue benci di jalanan selain abang-abang maxim yang nabrak knalpot gue sampai patah adalah abang-abang gojek yang bawa motor ngebut di tikungan kayak tolol anjing.
Sesudah
orang yang gue benci di jalan selain abang-abang maxim yang nabrak knalpot gue sampai patah adalah abang-abang gojek yang bawa motor ngebut di tikungan kayak tolol anjing.

d. *Stopwords Filtering*

Pada tahap ini, dilakukan penghapusan kata-kata yang dianggap tidak relevan untuk meningkatkan efisiensi pemrosesan data. Hanya kata-kata yang dianggap relevan dan bermakna penting

yang dipertahankan. Hasil dari proses penyaringan stopwords dapat dilihat dalam Tabel 6.

Tabel 6. Hasil Proses Stopwords Filtering

Sebelum
saya suka naik gojek karena nyaman dan cepat
Sesudah
suka naik gojek nyaman cepat

4. *Transformation (TF-IDF)*

TF-IDF adalah metode yang umum digunakan dalam pengolahan teks dan Information Retrieval untuk mengevaluasi pentingnya sebuah kata dalam sebuah dokumen dalam kumpulan dokumen. Metode ini mengkombinasikan dua konsep utama: Term Frequency (TF) dan Inverse Document Frequency (IDF):

a. *Term Frequency (TF)*

TF mengukur seberapa sering sebuah kata muncul dalam sebuah dokumen. Nilai TF untuk sebuah kata dalam sebuah dokumen dapat dihitung dengan cara yang sederhana, yaitu dengan membagi jumlah kemunculan kata tersebut dengan jumlah total kata dalam dokumen tersebut. TF meningkat seiring dengan peningkatan frekuensi kata dalam dokumen tersebut.

b. *Inverse Document Frequency (IDF)*

IDF mengukur seberapa jarang sebuah kata muncul di seluruh kumpulan dokumen. Kata-kata yang jarang muncul di banyak dokumen akan memiliki nilai IDF yang tinggi, sementara kata-kata yang sering muncul di banyak dokumen akan memiliki nilai IDF yang rendah. Nilai IDF untuk sebuah kata dapat dihitung dengan cara mengambil logaritma dari jumlah total dokumen dalam kumpulan dokumen dibagi dengan jumlah dokumen yang mengandung kata tersebut, kemudian hasilnya dimuluskan

TF-IDF adalah hasil perkalian antara Term Frequency (TF) dan Inverse Document Frequency (IDF) untuk setiap kata dalam dokumen. Dengan menggunakan metode ini, kata-kata yang muncul sering dalam dokumen tersebut namun jarang muncul dalam kumpulan dokumen secara keseluruhan akan memiliki bobot yang tinggi, menunjukkan pentingnya kata tersebut dalam dokumen tersebut. Dengan menerapkan metode TF-IDF, kita dapat mengubah representasi data teks menjadi vektor numerik yang menggambarkan tingkat pentingnya setiap kata dalam dokumen.

Metode ini diterapkan pada penelitian ini menggunakan TfidfVectorizer di dalam pemrograman python. TfidfVectorizer adalah alat dalam library scikit-learn yang mengubah sekelompok dokumen teks menjadi representasi numerik berdasarkan skema TF-IDF. Proses ini berperan mengubah teks menjadi vektor fitur TF-IDF yang dapat dimengerti oleh model Naive Bayes.

5. *Naive Bayes Classifier*

Pengklasifikasian menggunakan algoritma Naive Bayes dilakukan dengan bahasa pemrograman Python. Setelah itu, algoritma Naïve Bayes diaplikasikan pada data uji dan performa model dihitung. Rancangan pemodelan pada pemrograman Python untuk membuat model klasifikasi dimulai dengan inisialisasi model MultinomialNB dalam scikit-learn. MultinomialNB adalah implementasi dari algoritma *Naive Bayes* yang digunakan untuk klasifikasi data dengan representasi fitur yang bersifat multinomial, seperti vektor frekuensi kata-kata dalam data teks yang telah diubah menjadi TF-IDF. Saat menggunakan MultinomialNB, langkah pertama adalah inisialisasi model, di mana model Naive Bayes Multinomial disiapkan untuk belajar dari data pelatihan menggunakan metode fit. Proses pelatihan ini melibatkan perhitungan probabilitas kemunculan setiap kata dalam setiap kategori berdasarkan data yang disediakan. Setelah dilatih, model dapat digunakan untuk memprediksi label untuk data uji yang belum pernah dilihat sebelumnya oleh model dengan menggunakan metode predict.

Setelah pembentukan model selesai, model tersebut akan diimplementasikan pada beberapa skema split data yang berbeda sebagai perbandingan, yaitu dengan rasio 80:20, 75:25, dan 70:30. Selanjutnya, dilakukan klasifikasi pada data dengan model, kinerja model akan menghasilkan nilai yang akan dimasukkan ke dalam *confusion matrix*. Berikut ini adalah hasil *confusion matrix* untuk masing-masing rasio yang dapat dilihat pada Tabel 5, 6, dan 7.

Tabel 5 Confusion Matrix Split Data 80:20

Prediksi	**Aktual**	
	Positif	**Negatif**
Positif	13	2
Negatif	1	3

Tabel 6 Confusion Matrix Split Data 75:25

Prediksi	**Aktual**	
	Positif	**Negatif**
Positif	15	3
Negatif	1	4

Tabel 7 Confusion Matrix Split Data 70:30

Prediksi	**Aktual**	
	Positif	**Negatif**
Positif	16	9
Negatif	1	2

6. *Evaluation*

Setelah proses klasifikasi dengan Naïve Bayes dilakukan, tahap berikutnya adalah proses evaluasi yang bertujuan untuk mengetahui nilai performance matrix yang meliputi *accuracy*, *precision*, dan *recall* dari hasil confusion matrix yang didapatkan. Berikut adalah performance matrix dari masing-masing split data yang dapat dilihat pada Tabel 8, 9, dan 10.

Tabel 8. Performance Matrix Split Data 80:20

	Persentase
Accuracy	84,21%
Precision	86,67%
Recall	92,86%

Dari Tabel 8, nilai *accuracy*, *precision*, dan *recall* dapat dihitung secara manual seperti berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{13 + 3}{13 + 3 + 2 + 1} = 0{,}8421 \times 100\% = 84{,}21\%$$

$$Precision = \frac{TP}{TP + FP} = \frac{13}{13 + 2} = 0{,}8667 \times 100\% = 86{,}67\%$$

$$Recall = \frac{TP}{TP + FN} = \frac{13}{13 + 1} = 0.9375 \times 100\% = 92{,}86\%$$

Tabel 9. Performance Matrix Split Data 75:25

	Persentase
Accuracy	82,60%
Precision	83,33%
Recall	93,75%

Dari Tabel 9, nilai *accuracy*, *precision*, dan *recall* dapat dihitung secara manual seperti berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{15 + 4}{15 + 4 + 3 + 1} = 0{,}8260 \times 100\% = 82{,}60\%$$

$$Precision = \frac{TP}{TP + FP} = \frac{15}{15 + 3} = 0{,}8333 \times 100\% = 83{,}33\%$$

$$Recall = \frac{TP}{TP + FN} = \frac{15}{15 + 1} = 0.9375 \times 100\% = 93{,}75\%$$

Tabel 10. Performance Matrix Split Data 70:30

	Persentase
Accuracy	64.29%
Precision	64.00%
Recall	94.12%

Dari Tabel 10, nilai *accuracy*, *precision*, dan *recall* dapat dihitung secara manual seperti berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{16 + 2}{16 + 2 + 9 + 1} = 0{,}6429 \times 100\% = 64.29\%$$

$$Precision = \frac{TP}{TP + FP} = \frac{16}{16 + 9} = 0.64 \times 100\% = 64.00\%$$

$$Recall = \frac{TP}{TP + FN} = \frac{16}{16 + 1} = 0.9412 \times 100\% = 94.12\%$$

Setelah melakukan evaluasi performance matrix pada split data dengan rasio 70:30, 75:25, dan 80:20, hasil kinerja masing-masing rasio akan dikomparasi. Tujuannya adalah untuk mengidentifikasi rasio split data yang menghasilkan kinerja model terbaik. Hasil komparasi ini dapat dilihat pada Tabel 11.

Tabel 11. Komparasi Performance Matrix

	Accuracy	**Precision**	**Recall**
80:20	84,21%	86,67%	92,86%
75:25	82,60%	83,33%	93,75%
70:30	64.29%	64.00%	94.12%

Performa klasifikasi sentimen menggunakan algoritma Naive Bayes dengan rasio 80:20, di mana 80% data digunakan untuk pelatihan dan 20% untuk pengujian, menunjukkan kinerja yang lebih unggul dibandingkan dengan rasio 75:25 dan 70:30 secara keseluruhan.Pada rasio ini dihasilkan tingkat accuracy sebesar 84,21%, nilai precision sebesar 86,67%, dan nilai recall sebesar 92,86%.

Pembahasan

Efektivitas Crawling Data

Proses Crawling Data berhasil untuk mengumpulkan 1000 tweet dengan kata kunci "Gojek". Namun, hanya 92 tweet yang benar-benar dapat dikategorikan sebagai sentimen setelah filtering. Ini menunjukkan bahwa tidak semua tweet yang mengandung kata kunci tersebut mengekspresikan sentimen yang jelas terhadap Gojek. Hal ini mungkin disebabkan oleh banyaknya tweet yang hanya menyebutkan Gojek tanpa memberikan opini atau perasaan tertentu.

Distribusi Sentimen

Dari 92 tweet yang teridentifikasi memiliki sentimen, 56 tweet bersifat positif dan 36 tweet bersifat negatif. Ini menunjukkan bahwa sentimen positif terhadap Gojek lebih dominan dalam sampel yang dianalisis. Hal ini bisa menjadi indikator bahwa secara umum, pengguna Twitter memiliki pandangan yang lebih positif terhadap layanan Gojek.

Efektivitas Preprocessing Data

Tahapan preprocessing (case folding, cleansing, stemming, dan stopwords filtering) terbukti efektif dalam menyederhanakan dan membersihkan data. Proses ini sangat penting untuk meningkatkan akurasi analisis sentimen dengan menghilangkan noise dan menyeragamkan format data.

Performa Klasifikasi Naive Bayes

Hasil evaluasi menunjukkan bahwa model Naive Bayes memberikan performa terbaik pada rasio split data 80:20, dengan accuracy 84,21%, precision 86,67%, dan recall 92,86%. Ini menunjukkan bahwa model cukup baik dalam mengklasifikasikan sentimen, terutama dalam mengidentifikasi sentimen positif (recall tinggi).

Pengaruh Rasio Split Data

Perbandingan performa pada berbagai rasio split data (80:20, 75:25, 70:30) menunjukkan bahwa rasio 80:20 memberikan hasil terbaik. Ini mungkin disebabkan oleh keseimbangan yang baik

antara jumlah data training yang cukup besar untuk melatih model dengan baik, dan data testing yang cukup untuk menguji generalisasi model.

Implikasi untuk Gojek

Hasil analisis sentimen ini memberikan gambaran umum tentang persepsi publik terhadap Gojek di Twitter. Dominasi sentimen positif bisa menjadi indikator bahwa strategi dan layanan Gojek diterima dengan baik oleh sebagian besar pengguna. Namun, adanya sentimen negatif juga menunjukkan area-area yang mungkin perlu perbaikan atau perhatian lebih lanjut dari pihak Gojek.

Keterbatasan Analisis

Perlu diingat bahwa analisis ini terbatas pada data Twitter dan mungkin tidak merepresentasikan keseluruhan opini publik tentang Gojek. Faktor-faktor seperti bias sampling (pengguna Twitter mungkin tidak mewakili seluruh pengguna Gojek) dan keterbatasan dalam deteksi sarkasme atau konteks yang lebih kompleks juga perlu dipertimbangkan dalam interpretasi hasil.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada rekan-rekan kelompok: Saudara Adhi, saudara Daniel, dan saudara Revano yang telah membantu dalam proses penelitian dan penulisan jurnal ini sejak dari pengumpulan data hingga pengumpulan jurnal. Tidak lupa juga ucapan terima kasih terhaturkan kepada para pembimbing yang telah memberikan arahan masukan selama proses penelitian. Ucapan terima kasih juga ditujukan kepada pihak Twitter atau yang kini berubah nama menjadi X yang telah menyediakan platform untuk pengumpulan data yang sangat penting dalam penelitian ini. Terakhir, tidak lupa juga penulis mengucapkan terima kasih kepada keluarga dan teman-teman yang telah memberikan dukungan moral dan semangat selama proses penelitian dan penulisan jurnal ini. Akhir kata, penulis berharap hasil penelitian ini dapat memberikan kontribusi positif bagi pengembangan analisis sentimen di media sosial dan pemahaman persepsi publik terhadap layanan aplikasi digital di Indonesia, khususnya Gojek.

SIMPULAN

Penelitian ini bertujuan untuk menganalisis sentimen pengguna Twitter terhadap layanan Gojek dengan menggunakan algoritma Naive Bayes. Melalui serangkaian tahapan yang meliputi pengumpulan data tweet, pelabelan data, praproses data, transformasi data dengan metode TF-IDF, klasifikasi menggunakan Naive Bayes, evaluasi model, dan visualisasi hasil, penelitian ini berhasil memberikan gambaran yang jelas mengenai sentimen pengguna terhadap layanan Gojek. Hasil analisis menunjukkan bahwa algoritma Naive Bayes mampu mengklasifikasikan sentimen tweet dengan accuracy sebesar 84,21%, precision sebesar 86,67%, serta recall sebesar 92,86% menggunakan rasio split data 80:20. Penelitian ini membuka peluang untuk pengembangan lebih lanjut dalam beberapa aspek. Pertama, peningkatan kualitas data dapat dilakukan dengan mengumpulkan lebih banyak data untuk meningkatkan keakuratan model. Kedua, penelitian selanjutnya dapat mengkaji penggunaan algoritma klasifikasi lain seperti SVM, Random Forest, atau Deep Learning untuk membandingkan kinerja dan akurasi dalam analisis sentimen. Algoritma-algoritma ini mungkin menawarkan pendekatan yang berbeda yang bisa memberikan hasil yang lebih baik atau lebih efisien dalam kondisi tertentu.

DAFTAR PUSTAKA

Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Proceedings of LREC. 10. Diambil dari https://www.researchgate.net/publication/220746311_Twitter_as_a_Corpus_for_Sentiment_Analysis_and_Opinion_Mining

Al Haromainy, M. M., Prasetya, D. A., & Sari, A. P. (2023). Improving

Performance of RNN-Based Models With Genetic Algorithm Optimization For Time Series Data. Diambil dari https://journal.undiknas.ac.id/index.php/tiers/article/view/4326

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Diambil dari https://doi.org/10.1016/j.asej.2014.04.011

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., Brown, D., Id, L., & Barnes. (2019). Text Classification Algorithms: A Survey. Diambil dari http://dx.doi.org/10.3390/info10040150

Rani, S. & Kumar, P. (2017). A Sentiment Analysis System to Improve Teaching and Learning. Diambil dari http://dx.doi.org/10.1109/MC.2017.133

Sukardi, S. (2014). Metodologi Penelitian Pendidikan: Kompetensi dan Praktiknya. Bumi Aksara.

Hodge, V. J., & Austin, J. (2021). A Survey of Outlier Detection Methodologies. Artificial Intelligence Review. Diambil dari http://dx.doi.org/10.1023/B:AIRE.0000045502.10941.a9

Taneja, A. (2024). SENTIMENT ANALYSIS USING MACHINE LEARNING: A COMPREHENSIVE REVIEW. Diambil dari https://www.researchgate.net/publication/382182195_SENTIMENT_ANALYSIS_USING_MACHINE_LEARNING_A_COMPREHENSIVE_REVIEW

Alfarizi, M., Syafaah, L., & Lestandy, M. (2022). Emotional Text Classification Using TF-IDF (Term Frequency-Inverse Document Frequency) And LSTM (Long Short-Term Memory). Diambil dari http://dx.doi.org/10.30595/juita.v10i2.13262

Kuhn, M., & Johnson, K. (2019). Applied Predictive Modeling. Springer.