# Analisis Sentimen Terhadap Ulasan Aplikasi Gojek Menggunakan Naive Bayes Classifier dengan BoW

Wayan Farel Nickholas Sadewa[a1], Luh Gede Astuti[a2]

Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Udayana
Jalan Raya Kampus UNUD, Bukit Jimbaran, Kuta Selatan, Badung, Bali, Indonesia
[1]farel.harri@gmail.com
[2]lg.astuti@unud.ac.id

***Abstract***

*In this digital era, technology pervades every aspect of daily life, revolutionizing industries and interactions. Within this landscape, online-based transportation services have emerged as transformative solutions, notably exemplified by Gojek in Indonesia. This study delves into sentiment analysis of Gojek reviews using Multinomial Naive Bayes and Bag-of-Words extraction, aiming to gauge user perceptions and responses. Leveraging a dataset of 9.996 App reviews, the research undertakes comprehensive preprocessing, including case folding, filtering, tokenization, stopword removal, and stemming, followed by sentiment labeling. By employing Bag-of-Words feature extraction, textual data is converted into numerical vectors, enabling the application of the Multinomial Naive Bayes classification model. Evaluation metrics, derived from a confusion matrix, reveal an accuracy rate of 86.29%, with precision, recall, and F1-Score values of 86.94%, 86.41%, and 86.26% respectively. This study underscores the efficacy of the adapted Multinomial Naive Bayes model with Bag-of-Words feature extraction in discerning user sentiments towards Gojek, offering valuable insights for enhancing service applications in the digital realm.*

***Keywords:*** *Sentiment Analysis, Gojek, Multinomial Naive Bayes, BoW*

## 1. Pendahuluan

Pada era digital ini, teknologi sudah menjadi sebuah inti dari kehidupan sehari-hari. Hampir keseluruhan bidang menggunakan teknologi. Mulai dari bisnis, kesehatan, hiburan, komunikasi, produktivitas, maupun penyediaan jasa secara *online*. Semua hal tersebut telah disesuaikan dengan kemajuan teknologi yang digunakan untuk meningkatkan efisiensi, produktivitas, maupun kualitas dari layanan. Dari berbagai perubahan ini tidak hanya merubah cara kita bekerja dan juga belajar. Namun, telah mempengaruhi bagaimana cara kita untuk berinteraksi dengan lingkungan sekitar kita dan juga masyarakat secara keseluruhan.

Dalam konteks penyediaan jasa angkutan umum, Terdapat berbagai variasi kehadiran jasa transportasi yang berbasis aplikasi *online*. Dengan mudahnya akses yang cepat melalui perangkat seluler, masyarakat tidak hanya mendapatkan kemudahan dalam mencari dan menggunakan layanan transportasi yang ada[1]. Dengan adanya pergeseran dari sistem angkutan umum konvensional menjadi sistem angkutan umum berbasis aplikasi *online*, para pengusaha yang bekerja pada bidang transportasi online dapat melakukan adaptasi serta menghadirkan solusi yang sangat inovatif yang dapat menguntungkan baik untuk pengguna jasa maupun penyedia jasa.

Ada banyak aplikasi penyedia jasa angkutan umum yang telah tersebar hingga saat ini, aplikasi Gojek menjadi salah satu aplikasi penyedia jasa angkutan umum yang paling menonjol. Dengan berbagai fitur yang ditawarkan, seperti pemesanan ojek, taksi, antar barang, pemesanan makanan, dan masih banyak fitur lainnya, Gojek telah sukses menghadirkan solusi yang sangat baik bagi kebutuhan transportasi dan layanan sehari-hari masyarakat indonesia[2]. Keberhasilan Gojek telah menjangkau berbagai segmen pasar dan juga memberikan layanan yang handal

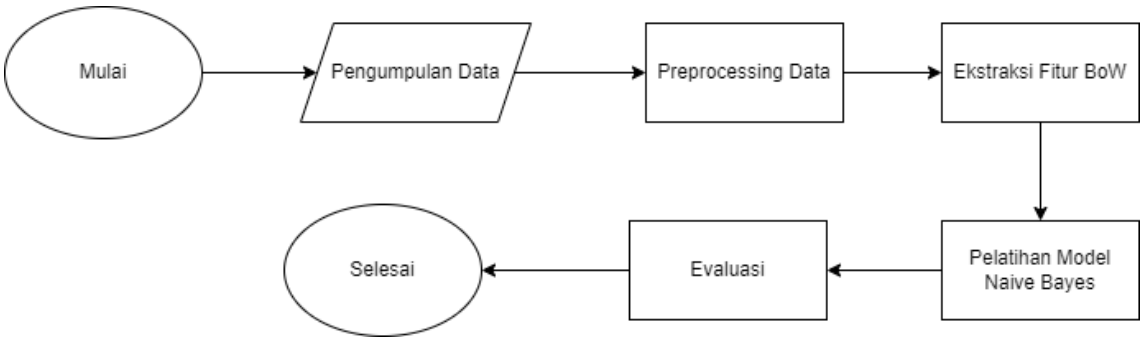
p-ISSN: 2986-3929
e-ISSN: 3032-1948

yang dapat mencerminkan kemampuan teknologi dalam memfasilitasi interaksi antara penyedia jasa maupun pengguna.

Untuk mengevaluasi adanya dampak dan respon terhadap layanan Gojek, dapat dilakukan analisis sentimen yang dimana berasal dari bidang penelitian *Natural Language Processing* (NLP) yang bertujuan untuk mengklasifikasikan data sentimen yang ada menjadi positif, negatif, maupun netral[3]. Dalam penelitian ini, akan digunakan metode Naive Bayes. Naive Bayes adalah metode klasifikasi yang didasarkan pada teorema Bayes. Dengan asumsi fitur yang digunakan untuk prediksi adalah independen satu sama lain[4]. Pada sentimen analisis, Naive Bayes dapat memberikan kinerja yang baik dalam pengolahan data teks yang besar dan juga kecepatan dalam pelatihan dan prediksi.

## 2. Metode Penelitian

### 2.1 Desain Penelitian




**Gambar 1.** Bagan Alur Penelitian

Pada permulaan penelitian ini, yang pertama kali dilakukan yaitu melakukan pengumpulan data ulasan aplikasi Gojek untuk melakukan sentimen analisis. Setelah itu data yang telah didapatkan akan melalui proses *preprocessing* data yang dimana data yang ada akan diolah hingga siap untuk digunakan pada model. Setelah itu dilakukannya *preprocessing* akan dilakukan ekstraksi fitur menggunakan Bag-of-Words (BoW). Kemudian melakukan pelatihan model naive bayes dengan ekstraksi fitur yang telah disiapkan sebelumnya. Tahapan terakhir yaitu evaluasi untuk mengetahui tingkat akurasi yang diberikan oleh model, dan Pada keseluruhan penelitian ini akan dilakukan menggunakan google colab dengan bahasa python

### 2.2 Pengumpulan Data

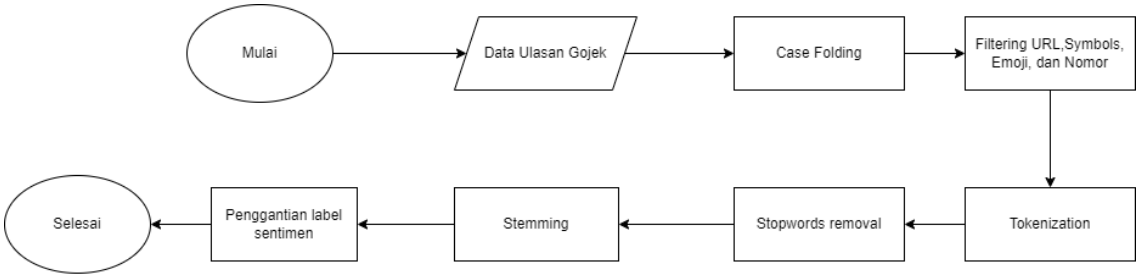
Teknik pengumpulan data yang akan digunakan pada penelitian ini menggunakan data sekunder yang diambil melalui *Kaggle*, yang dimana dicari dengan kata kunci "*reviews*". Dataset yang digunakan merupakan dataset review Gojek dari Google Playstore yang dimana terdapat sebanyak 9.996 data. Pada data *Kaggle* tidak terdapat label sentimen namun terdapat *score* yang dimana dapat diasumsikan score 1-3 merupakan sentimen negatif, dan *score* 4-5 merupakan sentimen positif. Namun yang akan digunakan hanya sentimen positif dan juga netral.

**Tabel 1.** Tabel Gambaran Dataset

| data_done | sentimen |
|---|---|
| rezpon untuk pelayanan go food lemot | negatif |
| Gojek | positif |
| Do blokir terus aja gopaylaternya gpp biar saya ga bisa jajan dan pake lagi 🙏 | negatif |

| data_done | sentimen |
|---|---|
| Tolong admin udah lihat yg viral sekarang gak? Byk oknum ojol yang melanggar lalu lintas seperti law... | negatif |
| Selalu pakai aplikasi gojek dong.... | positif |

### 2.3 Text Preprocessing




**Gambar 2.** Bagan Alur Tahap Preprocessing

Pada tahap preprocessing data yang pertama kali dilakukan yaitu *case folding*, yang dimana mengubah semua huruf menjadi huruf kecil, setelah itu dilakukan filtering untuk menghilangkan URL, simbol, emoji, dan juga nomor yang ada pada dataset. Setelah itu dilakukan *tokenization* untuk mengubah tiap kata menjadi token-token. Dan dilakukan *stopwords removal* untuk menghapus kata-kata yang sekiranya tidak terkait dengan data dan tidak memberikan makna tambahan pada dataset. Yang terakhir yaitu tahap *stemming* untuk mengubah kata-kata yang ada menjadi kata dasar. Setelah tahap *Stemming score* juga akan diubah menjadi kolom sentimen yang akan diubah menjadi sentimen negatif (1-2), sentimen positif (4-5).

### 2.4 Ekstraksi Fitur Bag-of-Words

*Bag-of-Words* (BoW) merupakan ekstraksi yang mengubah dokumen atau teks menjadi vektor berdasarkan kemunculan kata-kata di dalamnya, dengan setiap kata dianggap sebagai fitur. Vektor yang dihasilkan mencerminkan frekuensi kemunculan kata-kata tersebut dalam teks, tanpa memperhatikan urutan kata-kata. BoW umumnya digunakan dalam analisis teks, termasuk analisis sentimen menggunakan *machine learning* untuk menghasilkan fitur-fitur yang digunakan dalam proses klasifikasi [5].

### 2.5 Naive Bayes

Multinomial Naive Bayes merupakan salah satu model Naive Bayes yang dimana setiap data yang ada harus dilabeli sebelum melakukan *training*[6]. Multinomial Naive Bayes adalah algoritma dengan waktu pemrosesan yang cepat, implementasi yang mudah, dan struktur yang sederhana. Dengan kemampuannya menghitung probabilitas kelas berdasarkan atribut, algoritma ini dapat mengidentifikasi kelas dengan probabilitas tertinggi. Pendekatan klasifikasi ini didasarkan pada asumsi bahwa setiap atribut dalam data bersifat eksklusif, mempermudah proses analisis dan interpretasi data[7].

$$P(C) = \frac{count(c)+K}{N+K.\ |classes|} \quad (1)$$

Yang dimana $P$ merupakan probabilitas yang ada pada variabel $c$ , $count(c)$ yang adalah kemunculan dari sampel $c$. $K$ yang adalah nilai dari parameter, $N$ yang adalah jumlah total kejadian pada sampel $c$ , dan $|classes|$ yang merupakan jumlah kelas yang ada pada sampel. Setelah itu dilakukan penghitungan nilai atribut, yang dimana jumlah atribut telah diberikan kelas kemudian ditambah satu dan dibagi dengan jumlah kelas tertentu yang ditambah dengan hasil

kali antara jumlah atribut dikali satu. Kemudian setelah terdapat persamaan untuk menghitung nilai $p(w|c)$ dengan *laplace smoothing*.

$$P(W|C) = \frac{count+K}{count(C)\ |v|.1} \quad (2)$$

dengan penjelasan bahwa $K$ adalah nilai parameter, $count(C)$ adalah jumlah kelas yang muncul pada sampel c, dan $|v|$ yang merupakan jumlah atribut yang terdapat pada sampel.

Pada penelitian ini, model Multinomial Naive Bayes ini akan menggunakan dataset yang sebelumnya telah dipreprocessing yang kemudian akan dilakukan ekstraksi fitur BoW untuk memperoleh nilai *accuracy*, *precision*, *recall*, dan F1-*Score* terbaik.

### 2.6 Confusion Matrix

Metode evaluasi yang akan diterapkan dalam penelitian ini merupakan *confusion matrix* yang merupakan metode untuk menguji suatu objek serta mampu membedakan antara kebenaran dan kesalahan[6]. Pada *confusion matrix* akan dilakukan 4 buah perhitungan, yaitu perhitungan *Accuracy*, *Precision*, *Recall*, dan juga F1-*Score*. Berikut merupakan rumus untuk melakukan perhitungan *confusion matrix*.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1 - Score = \frac{2.Precision.Recall}{Precision+Recall} \quad (6)$$

## 3. Hasil dan Pembahasan

### 3.1 Preprocessing Data

Berdasarkan hasil *preprocessing* data yang dimana terdiri dari 4997 data positif, dan 4997 data negatif. *Case Folding* dilakukan untuk mengkonversi semua huruf dalam dataset menjadi huruf kecil. Kemudian terdapat penghapusan URL, Simbol, Emoji, dan Nomor pada dataset. Setelah itu dilakukan penghapusan *stopwords* dengan library Sastrawi untuk menghapus *stopwords* berbahasa Indonesia. Dan yang terakhir yaitu tahap stemming yang dilakukan dengan library Sastrawi untuk mengubah kata-kata yang ada menjadi bentuk dasarnya. Adapun contoh data yang telah melewati tahap *preprocessing* terdapat pada tabel 2.

**Tabel 2.** Contoh Data Preprocessing

| | **data_done** | **sentimen** |
|---|---|---|
| **Data Awal** | Gak bisa login ke akun lama. | negatif |
| **Case Folding** | gak bisa login ke akun lama. | negatif |
| **Filtering URL, Simbol, Emoji, Nomor** | gak bisa login ke akun lama | negatif |
| **Tokenization** | [gak, bisa, login, ke, akun, lama] | negatif |
| **Stopwords Removal** | [gak, login, akun, lama] | negatif |
| **Stemming** | [gak, login, akun, lama] | negatif |

p-ISSN: 2986-3929
e-ISSN: 3032-1948

### 3.2 Perhitungan Ekstraksi Fitur Bag-of-Words

Setelah data melalui tahap *preprocessing*, langkah berikutnya merupakan perhitungan ekstraksi fitur BoW yang dimana data akan diubah menjadi representasi numerik dan akan direpresentasikan sebagai vektor BoW. Ekstraksi fitur ini dilakukan menggunakan *library scikit-learn*. *CountVectorizer* untuk membantu dalam mengubah teks yang ada menjadi vektor BoW. *CountVectorizer* akan menghitung seberapa sering sebuah kata muncul yang ada pada dataset. Kata dalam dokumen kemudian membangun vektor BoW berdasarkan kemunculan kata-kata yang ada. Hasil dari ekstraksi fitur kemudian akan digunakan dalam proses pelatihan model klasifikasi Naive Bayes.

```
[26] vectorizer = CountVectorizer()
     x_train_vectorized = vectorizer.fit_transform(x_train)
     x_test_vectorized = vectorizer.transform(x_test)
```

**Gambar 3.** Implementasi Ekstraksi Fitur Bag-of-Words

### 3.3 Pemodelan Naive Bayes

Setelah ekstraksi fitur BoW, langkah berikutnya merupakan pemodelan metode klasifikasi Naive Bayes. Metode ini cocok digunakan untuk klasifikasi teks dengan representasi vektor BoW. Sebelum melakukan pemodelan, dilakukan *splitting* data menjadi data latih dan juga data uji dengan perbandingan data uji sebesar 20%, dan data latih sebesar 80% yang dapat dilihat pada gambar 4.

```
[6] x_train, x_test, y_train, y_test = train_test_split(data['data_done'], data['sentimen'], test_size=0.2, random_state=0)

[7] print(f'jumlah data latih:  {len(x_train)}')
    print(f'jumlah data uji:  {len(x_test)}')

    jumlah data latih:  7995
    jumlah data uji:  1999
```

**Gambar 4.** Splitting Data Latih dan Data Uji

Dilakukan pemodelan data Multinomial Naive Bayes dengan bantuan *library sci-kit learn*. Setelah itu dilakukan tuning hyperparameter alpha sebesar 0.2 yang digunakan untuk menangani kasus di fitur yang tidak muncul pada data pelatihan. Dan kemudian model dilatih dengan data latih yang telah di*split* sebelumnya.

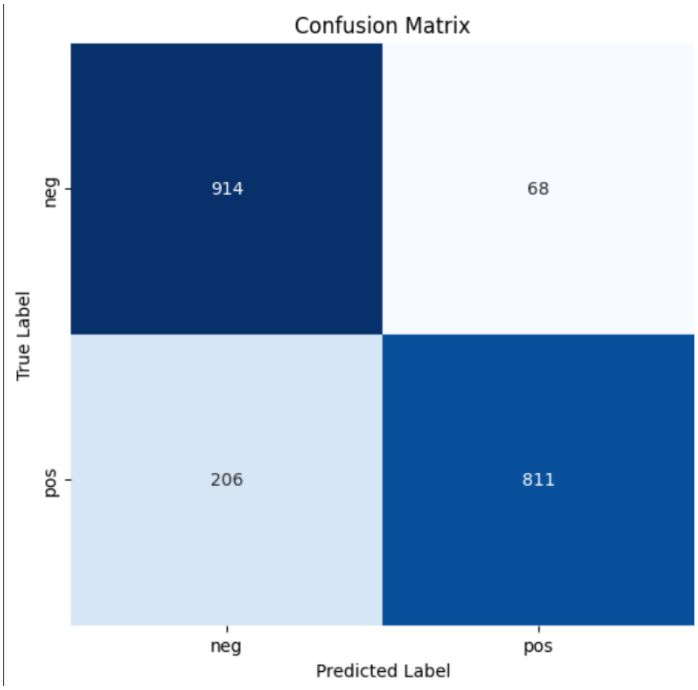
```
0s [80] model = MultinomialNB(alpha=.2)
        model.fit(x_train_vectorized, y_train)

        MultinomialNB
        MultinomialNB(alpha=0.2)
```

**Gambar 5.** Pelatihan model Multinomial Naive Bayes

Setelah melatih model Multinomial Naive Bayes, langkah selanjutnya yaitu menguji model menggunakan data uji yang telah diproses dengan ekstraksi fitur BoW sebelumnya. Setelah melakukan prediksi terhadap data uji, hasil prediksi akan dibandingkan dengan label aslinya dari data uji yang akan menghasilkan *confusion matrix*. *Confusion matrix* adalah tabel uji yang

p-ISSN: 2986-3929
e-ISSN: 3032-1948

menyajikan jumlah prediksi benar dan salah yang telah dibuat oleh model klasifikasi pada setiap kelas.




**Gambar 6.** Confusion Matrix Bag-of-Words

Berdasarkan *confusion matrix* pada gambar 6, model berhasil memprediksi sebanyak 914 sentimen negatif, namun terdapat 68 data bersentimen negatif yang diprediksi positif. Kemudian model berhasil memprediksi sebanyak 811 sentimen positif, namun terdapat 206 sentimen positif yang diprediksi negatif oleh model. Berdasarkan confusion matrix diatas didapatkan nilai *Accuracy*, *Precision*, F1-*Score* pada model Multinomial Naive Bayes..

Hasil Klasifikasi:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| negatif | 0.8161 | 0.9308 | 0.8696 | 982 |
| positif | 0.9226 | 0.7974 | 0.8555 | 1017 |
| accuracy | | | 0.8629 | 1999 |
| macro avg | 0.8694 | 0.8641 | 0.8626 | 1999 |
| weighted avg | 0.8703 | 0.8629 | 0.8624 | 1999 |

**Gambar 7.** Hasil Evaluasi

Dari hasil akurasi yang didapatkan model Multinomial Naive Bayes yang telah dilatih dengan ekstraksi fitur Bag-of-Words, didapatkan nilai akurasi sebesar 86.29%, nilai precision 86.94%, nilai recall 86.41%, dan juga nilai F1-Score 86.26%

## 4. Kesimpulan

Dari hasil penelitian ini, dapat ditarik kesimpulan bahwa model Multinomial Naive Bayes yang telah disesuaikan dengan dataset ulasan aplikasi Gojek dari Google Playstore dan ditunjang dengan ekstraksi fitur Bag-of-Words serta *tuning hyperparameter alpha* pada tingkat 0.2, berhasil

mengklasifikasikan sentimen ulasan dengan akurasi yang tinggi, mencapai 86.29%. Evaluasi lanjutan juga menunjukkan nilai *precision*, *recall*, dan F1-*Score* yang memuaskan, yaitu masing-masing sebesar 86.94%, 86.41%, dan 86.26%. Penelitian ini menegaskan bahwa model dapat menjadi instrumen yang efektif dalam memahami tanggapan pengguna terhadap layanan aplikasi Gojek.

**Daftar Pustaka**

[1] A. Luwihono, A. F. Peranginangin, Bahtiar, and A. Makanuay, "Peningkatan Moda Transportasi Indonesia Era Digital Berbasiskan Sistem Cerdas," *ATDS SAINTECH-Journal Eng.*, vol. 1, no. 2, pp. 55–63, 2020, [Online]. Available: https://ojs.atds.ac.id/index.php/atdssaintech/article/view/14

[2] F. D. Amajida, "Kreativitas Digital Dalam Masyarakat Risiko Perkotaan: Studi Tentang Ojek Online 'Gojek' Di Jakarta," *Informasi*, vol. 46, no. 1, p. 115, 2016, doi: 10.21831/informasi.v46i1.9657.

[3] A. Rajput, "Natural language processing, sentiment analysis, and clinical analytics," *Innov. Heal. Informatics A Smart Healthc. Prim.*, pp. 79–97, 2019, doi: 10.1016/B978-0-12-819043-2.00003-4.

[4] J. N. Sulzmann, J. Fürnkranz, and E. Hüllermeier, "On pairwise naive bayes classifiers," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4701 LNAI, pp. 371–381, 2007, doi: 10.1007/978-3-540-74958-5_35.

[5] H. Utama and A. Masruro, "Analisis Sentimen pada Twitter menggunakan Word Embedding dengan Pendekatan Word2Vec," *J. Sist. Cerdas*, vol. 5, no. 2, pp. 128–134, 2022, doi: 10.37396/jsc.v5i2.242.

[6] F. J. Damanik and D. B. Setyohadi, "Analysis of public sentiment about COVID-19 in Indonesia on Twitter using multinomial naive bayes and support vector machine," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 704, no. 1, 2021, doi: 10.1088/1755-1315/704/1/012027.

[7] A. A. Ajhari, "The Comparison of Sentiment Analysis of Moon Knight Movie Reviews between Multinomial Naive Bayes and Support Vector Machine," *Appl. Inf. Syst. Manag.*, vol. 6, no. 1, pp. 13–20, 2023, doi: 10.15408/aism.v6i1.26045.

p-ISSN: 2986-3929
e-ISSN: 3032-1948

Halaman ini sengaja dibiarkan kosong