

Exploratory Data Analysis Pada Dataset Ulasan Pelanggan Mc'Donald di Australia

<http://dx.doi.org/10.28932/jste.vXiX.X>

Received: 28 Agustus 2025 | Revised: 27 September 2025 | Accepted: 29 September 2025

Creative Commons License 4.0 (CC BY – NC)



William Wijaya^{✉#1}, Setia Budi^{*2}

*#Program Studi Teknik Informatika, Fakultas Teknologi dan Rekayasa Cerdas, Universitas Kristen Maranatha
Jl. Prof. drg. Surya Sumantri No.65, Bandung, Jawa Barat 40164, Indonesia*

¹2172011@maranatha.ac.id

**Program Studi Magister ilmu Komputer, Fakultas Teknologi dan Rekayasa Cerdas, Universitas Kristen Maranatha
Jl. Prof. drg. Surya Sumantri No.65, Bandung, Jawa Barat 40164, Indonesia*

²setia.budi@it.maranatha.edu

✉Corresponding author: 2172011@maranatha.ac.id

How to cite this article:

W. Wijaya, S. Budi, “Exploratory Data Analysis Pada Dataset Ulasan Pelanggan Mc'Donald di Australia,” *Journal of Smart Technology and Engineering*, vol. 1, no. 1, 01–XX. pp. 59-71, <https://doi.org/10.28932/jts.vXXiXX.XXXX>

Abstrak — Penelitian ini, yang berjudul “Analisis Ulasan Pelanggan McDonald's di Australia”, bertujuan untuk meningkatkan kualitas layanan, yang merupakan aspek penting bagi restoran cepat saji dalam menjaga kepuasan pelanggan. Studi ini menganalisis ulasan pelanggan McDonald's di Australia untuk mengidentifikasi faktor-faktor utama yang memengaruhi pengalaman pelanggan. Data yang digunakan berasal dari dataset ulasan pelanggan McDonald's yang tersedia di Kaggle. Metode yang digunakan meliputi analisis sentimen berbasis Natural Language Processing (NLP) dan teknik analisis data eksploratif untuk mengidentifikasi tren umum dalam ulasan tersebut. Hasil penelitian menunjukkan bahwa faktor-faktor seperti kualitas makanan, kecepatan layanan, kebersihan restoran, dan interaksi dengan staf memiliki pengaruh signifikan terhadap sentimen pelanggan. Temuan ini dapat menjadi dasar bagi McDonald's untuk memperbaiki aspek layanan yang masih kurang sekaligus memperkuat elemen-elemen yang telah mendapatkan umpan balik positif dari pelanggan. Dengan memanfaatkan analisis data berbasis NLP, restoran dapat lebih memahami preferensi pelanggan dan meningkatkan strategi layanan mereka untuk meningkatkan loyalitas pelanggan.

Kata Kunci— Analisis Sentimen; Kualitas Layanan; McDonald's; *Natural Language Processing* (NLP); Ulasan Pelanggan.

Exploratory Data Analysis on McDonald's Customer Review Dataset in Australia

Abstract — This research, titled “Analysis of Customer Reviews of McDonald's in Australia” aims to enhance service quality, which is a crucial aspect for fast-food restaurants in maintaining customer satisfaction. This study analyzes McDonald's customer reviews in Australia to identify key factors influencing customer experience. The data used comes from the McDonald's customer review dataset available on Kaggle. The methods employed include sentiment analysis based on Natural Language Processing (NLP) and exploratory data analysis techniques to identify common trends in the reviews. The results indicate that factors such as food quality, service speed, restaurant cleanliness, and staff interaction significantly impact customer sentiment. These findings can serve as a foundation for McDonald's to improve underperforming service aspects while reinforcing elements that have received positive customer feedback. By leveraging NLP-based data analysis, restaurants can better understand customer preferences and enhance their service strategies to boost customer loyalty.

Keywords— *Customer Reviews; McDonald's; Sentiment Analysis; Service Quality; Natural Language Processing* (NLP).

I. PENDAHULUAN

Restoran cepat saji merupakan bagian dari industri jasa makanan yang sangat populer dan diminati oleh masyarakat modern. Perkembangan teknologi dan tren yang dinamis membuat industri ini semakin berkembang pesat, terutama dengan adanya media sosial yang dapat menyebarkan informasi dengan sangat cepat. Salah satu perubahan signifikan dalam industri ini adalah kemudahan akses pemesanan makanan secara daring melalui aplikasi atau *platform* digital. Fasilitas ini memungkinkan pelanggan untuk memesan makanan dengan praktis kapan saja dan di mana saja.

Selain memudahkan pemesanan, perkembangan teknologi juga memberikan pelanggan kesempatan untuk membagikan ulasan atau *review* mengenai pengalaman mereka dalam mengonsumsi makanan di restoran cepat saji. Ulasan ini dapat diunggah melalui media sosial maupun *platform* pemesanan makanan, sehingga informasi mengenai kualitas makanan dan layanan dapat tersebar luas dalam waktu singkat. Dengan semakin banyaknya ulasan pelanggan yang tersedia secara daring, restoran cepat saji memiliki peluang untuk memahami kebutuhan dan ekspektasi pelanggan dengan lebih baik.

Seiring dengan perubahan tren yang terus berkembang, analisis terhadap ulasan pelanggan menjadi aspek yang sangat penting bagi restoran cepat saji. *Review* yang diberikan pelanggan dapat memberikan wawasan mengenai tingkat kepuasan pelanggan terhadap kualitas makanan, kecepatan pelayanan, kebersihan restoran, serta interaksi dengan staf. Dengan melakukan analisis terhadap data ulasan pelanggan, restoran dapat mengidentifikasi aspek yang perlu ditingkatkan untuk memberikan layanan yang lebih baik serta mempertahankan loyalitas pelanggan.

Penelitian ini memiliki keunikan karena menggunakan pendekatan *multi-model* untuk analisis sentimen terhadap ulasan pelanggan *McDonald's* di Australia, dengan Fokus pada konteks linguistik khas Australia melalui gaya penyampaian dan kata-kata yang sering digunakan dalam review konsumen, tanpa membahas penggunaan *Australian* slang secara khusus. Selain itu, komentar dalam *dataset* bersifat independen dan tidak menunjukkan interaksi sosial antarpengguna, sehingga memungkinkan penilaian sentimen secara objektif per ulasan. *McDonald's* dipilih sebagai objek studi karena merupakan salah satu jaringan restoran cepat saji terbesar dengan kehadiran kuat di Australia (*big chain brand*), menjadikan analisis terhadap layanan dan kepuasan pelanggannya relevan dan berdampak luas.

II. KAJIAN TEORI

A. Exploratory Data Analysis

Analisis setimen merupakan sebuah bidang ilmu dari pemrosesan bahasa alami (*Natural Language Processing*) yang biasanya digunakan pada pemrosesan data untuk mengidentifikasi pendapat berdasarkan suatu topik yang berguna untuk menganalisis, memproses, serta mengekstraksi data tekstual mengenai entitas seperti layanan, produk, acara, dan topik tertentu. Analisis sentiment memiliki tugas utama yaitu mengklasifikasikan teks ke dalam kalimat atau dokumen yang kemudian menentukan penilaian dari kalimat tersebut apakah positif, negatif, atau netral. Selain memberikan nilai, kalimat yang telah diklasifikasikan dapat mengungkapkan perasaan emosional yang berupa perasaan sedih, senang, tidak senang, dan mengungkapkan pendapat mengenai sebuah barang atau *brand*, menganalisis suasana hati, pendapat, dan perasaan pengguna yang memberikan ulasan [1].

Seiring berkembangnya teknologi analisis sentimen, banyak studi terbaru yang menggabungkan metode machine learning dan deep learning dalam pengolahan data opini publik. Salah satu pendekatan yang umum digunakan adalah menerapkan model pre-trained seperti BERT, LSTM, dan BiGRU untuk mengklasifikasikan ulasan atau opini dalam skala besar dengan akurasi yang tinggi. Studi oleh Atlas et al. (2025) menunjukkan bahwa pendekatan ini mampu mencapai akurasi lebih dari 91% dalam analisis sentimen produk dengan hanya dua label (positif dan negatif) [2]. Di sisi lain, Kayed et al. (2023) menekankan pentingnya pemilihan fitur dan arsitektur model yang tepat untuk meningkatkan performa dalam tugas klasifikasi sentimen [3]. Hal ini memperlihatkan bahwa analisis sentimen bukan hanya tentang menentukan polaritas suatu kalimat, melainkan juga mengintegrasikan konteks, emosi, dan tujuan dari ekspresi yang diberikan.

B. Python

Python adalah salah satu bahasa komputer tingkat tinggi yang dapat digunakan untuk banyak aplikasi. Bahasa pemrograman Python memiliki keunggulan yaitu mudah digunakan untuk pemula dan cukup mudah untuk dipelajari. Selain itu, bahasa pemrograman Python sangat fleksibel karena tidak memiliki aturan ketat untuk membuat atau menggunakan fitur yang telah disediakan. Kelebihan lain yang dimiliki bahasa pemrograman Python adalah memiliki komunitas pengguna yang mendukung serta banyak paket yang dapat dipasang dan digunakan dengan mudah. Python merupakan bahasa pemrograman *open source* (sumber terbuka) yang dapat memudahkan seluruh pemakai dalam mempelajari bahasa pemrograman tersebut. Selain dari paket yang disediakan oleh bahasa pemrograman Python ataupun paket yang dikembangkan oleh pengguna lain dapat dipelajari serta digunakan secara gratis [4].

Bahasa pemrograman Python sering digunakan dalam menganalisis data karena Python memiliki banyak fitur serta memiliki pustaka yang luas, dan dapat digunakan untuk berbagai aplikasi seperti pemrograman statistik. Keunggulan lain dari Python adalah memiliki fitur untuk memvisualisasikan data secara interaktif seperti membuat bagan atau grafik [5].

C. *Exploratory Data Analysis*

Exploratory Data Analysis (EDA) adalah proses menganalisis, menampilkan data untuk mempermudah pemahaman dan wawasan dari data yang sedang dianalisa. Tujuan utama EDA adalah menganalisa serta memaksimalkan pemahaman mengenai kumpulan data serta menentukan faktor-faktor yang tidak terduga dan mempertimbangkan dampak dari faktor-faktor tersebut [6].

D. *Pandas*

Pandas adalah salah satu pustaka dalam Python yang banyak digunakan untuk manipulasi, analisis, dan pemrosesan data berbentuk tabel atau data terstruktur lainnya. Pustaka ini menyediakan berbagai fungsi untuk membaca, menulis, membersihkan, mengelola, dan menganalisis data dalam format seperti CSV, Excel, SQL, dan JSON [7]. Pandas sangat berguna dalam proses analisis data karena menyediakan struktur data yang fleksibel, seperti DataFrame dan Series, yang memungkinkan pengguna untuk mengelola data dalam berbagai bentuk dengan efisien.

Dalam analisis data, Pandas digunakan untuk berbagai operasi seperti filtering, grouping, merging, serta visualisasi awal terhadap dataset. Keunggulan Pandas dibandingkan pustaka lain adalah kemampuannya dalam menangani data yang besar dan kompleks dengan kinerja yang optimal [8]. Selain itu, Pandas sering digunakan dalam berbagai bidang seperti data science, machine learning, serta analisis bisnis karena kemampuannya dalam mengolah data dengan cepat dan mudah [9].

E. *Pre-Trained Model*

Pre-trained model adalah model pembelajaran mesin yang telah dilatih sebelumnya pada sejumlah besar data dan dapat langsung digunakan atau disesuaikan (fine-tuned) untuk tugas spesifik seperti analisis sentimen, klasifikasi teks, dan pemahaman bahasa alami. Penggunaan pre-trained model sangat menguntungkan dalam pengolahan bahasa alami karena menghemat waktu pelatihan, meningkatkan akurasi, dan memungkinkan penggunaan arsitektur yang kompleks tanpa memulai dari awal. Beberapa contoh pre-trained model yang populer dalam analisis sentimen termasuk BERT, DistilBERT, dan RoBERTa. Model-model ini telah terlatih pada korpus besar seperti Wikipedia atau BookCorpus dan memiliki kemampuan untuk menangkap representasi semantik dari teks [10].

F. *Natural Language Toolkit (NLTK)*

Natural Language Toolkit (NLTK) adalah sebuah pustaka dalam Python yang digunakan untuk pemrosesan bahasa alami. Pustaka NLTK menyediakan berbagai alat penting untuk melakukan operasi dasar seperti tokenisasi (pemecahan teks menjadi kata atau kalimat), stemming (mengubah kata menjadi bentuk dasar), tagging (penandaan bagian dari kalimat), parsing (analisis struktur kalimat), dan analisis semantik. Modul-modul tersebut digunakan untuk pre-processing teks sebelum melakukan analisis sentimen. Selain menyediakan modul yang dapat digunakan untuk mempersiapkan teks sebelum dilakukannya analisis sentimen, NLTK juga menyediakan pre-trained model yang dapat langsung digunakan untuk melakukan analisis sentimen. Salah satu contoh pre-trained model yang populer digunakan untuk analisis sentimen dari NLTK adalah VADER (Valence Aware Dictionary and sEntiment Reasoner) [11].

G. *Scikit-learn (Sklearn)*

Scikit-learn (sklearn) merupakan sebuah pustaka dalam Python yang menyediakan alat untuk *machine learning* dan analisis data, seperti klasifikasi, regresi, clustering, dan reduksi dimensi. Sklearn dalam analisis sentimen digunakan untuk membangun model klasifikasi seperti Naive Bayes, Support Vector Machine (SVM), dan Random Forest. Pustaka ini mendukung pipeline yang memudahkan preprocessing data, ekstraksi fitur, dan evaluasi model. Sklearn juga dapat digunakan bersama dengan pustaka NLP lainnya seperti NLTK atau TextBlob untuk melakukan analisis sentimen yang lebih komprehensif [12].

H. *VADER*

VADER (Valence Aware Dictionary and sEntiment Reasoner) adalah pre-trained model untuk analisis sentimen berbasis leksikon yang dirancang untuk media sosial dan teks informal. VADER memberikan nilai terhadap sebuah teks dengan nilai sentimen positif, negatif, dan netral [13].

I. *TextBlob*

TextBlob adalah pustaka Python yang digunakan untuk pemrosesan bahasa alami (NLP) yang menyediakan antarmuka sederhana seperti tokenisasi, tagging, ekstraksi frasa nominal, analisis sintaks, serta analisis sentimen. TextBlob menggunakan pendekatan berbasis leksikon dan pembelajaran mesin sederhana untuk mengklasifikasikan teks dengan nilai sentimen, positif, negatif, atau netral, dan memberikan skor polaritas. TextBlob memiliki keunggulan dalam kemudahan penggunaan serta memiliki integrasi yang baik dengan pustaka lain seperti NLTK dan scikit-learn (sklearn) [14].

J. *DistilBERT*

DistilBERT merupakan versi ringan dari BERT (Bidirectional Encoder Representations from Transformers) yang dikembangkan oleh Hugging Face dengan tujuan mengurangi ukuran model dan waktu komputasi tanpa kehilangan performa model yang signifikan. DistilBERT mempertahankan 97% dari akurasi BERT pada berbagai tugas NLP, termasuk analisis

sentimen, namun hanya menggunakan 60% jumlah parameter dan 40% waktu inferensi. Model ini sangat cocok untuk diterapkan dalam aplikasi yang membutuhkan efisiensi tinggi seperti di perangkat edge atau dengan sumber daya terbatas. Dalam analisis sentimen, DistilBERT mampu memahami konteks kata dan struktur kalimat secara mendalam berkat arsitektur transformer yang digunakan [15].

K. Flair

Flair adalah pustaka NLP berbasis Python yang dikembangkan oleh Zalando Research. Flair mendukung berbagai model pre-trained untuk berbagai tugas NLP seperti named entity recognition (NER), part-of-speech tagging, dan analisis sentimen. Salah satu fitur unggulannya adalah kemampuan untuk melakukan word embeddings kontekstual menggunakan model seperti BERT, GloVe, dan ELMo. Dalam analisis sentimen, Flair dapat digunakan untuk membangun model yang sangat akurat karena kemampuannya dalam menangkap makna kontekstual kata dalam kalimat. Flair juga menyediakan pre-trained sentiment classifier yang dilatih pada berbagai jenis data teks, termasuk review produk dan media sosial [16].

L. WordCloud

WordCloud adalah teknik visualisasi teks yang digunakan untuk menampilkan kata-kata yang paling sering muncul dalam kumpulan dokumen atau dataset. Dalam konteks analisis sentimen, WordCloud membantu dalam eksplorasi data dengan memvisualisasikan kata-kata yang sering digunakan oleh pengguna dalam ulasan atau komentar. Kata-kata dengan frekuensi tertinggi ditampilkan dalam ukuran yang lebih besar, memberikan gambaran cepat tentang topik atau opini yang dominan. Visualisasi ini sering digunakan pada tahap awal analisis untuk memahami karakteristik umum dari data teks [17].

M. Seaborn

Seaborn adalah pustaka visualisasi data berbasis Python yang dikembangkan di atas Matplotlib dan dirancang untuk membuat grafik statistik yang informatif dan menarik [18]. Pustaka ini menawarkan berbagai fitur seperti plot distribusi, hubungan antar variabel, serta visualisasi data kategori.

Keunggulan Seaborn terletak pada kemampuannya dalam menangani data yang kompleks serta integrasi yang baik dengan Pandas, sehingga memudahkan analisis eksploratif dalam penelitian data. Seaborn juga mendukung pengaturan tema otomatis serta palet warna yang dapat meningkatkan interpretasi data [19].

N. Matplotlib

Matplotlib adalah pustaka visualisasi data yang memungkinkan pengguna untuk membuat berbagai jenis grafik, termasuk line plot, bar chart, histogram, dan scatter plot [20]. Matplotlib sering digunakan dalam analisis data karena fleksibel dalam menyesuaikan tampilan visualisasi sesuai kebutuhan dan dapat mengontrol elemen grafik seperti warna, label, skala, serta anotasi yang fleksibel sehingga mudah dan praktis yang dapat digunakan untuk visualisasi data dalam berbagai situasi [21].

O. Justifikasi Penggunaan Dua Kelas Sentimen

Dalam penelitian ini, analisis sentimen dilakukan dengan menggunakan dua kelas saja, yaitu positif dan negatif, tanpa menyertakan kelas netral. Keputusan ini diambil berdasarkan beberapa pertimbangan teknis dan didukung oleh referensi penelitian terkini.

Pertama, dalam konteks ulasan konsumen, ekspresi netral cenderung bersifat ambigu dan tidak memberikan informasi yang cukup untuk pengambilan keputusan yang bermakna. Oleh karena itu, penghilangan kelas netral bertujuan untuk meningkatkan ketajaman analisis dan fokus pada ulasan yang memiliki makna opini yang kuat. Kedua, studi oleh Chatterjee et al. (2025) menunjukkan bahwa pendekatan dua kelas menghasilkan akurasi yang lebih tinggi dibandingkan dengan tiga kelas, terutama saat data netral memiliki jumlah terbatas atau distribusinya tidak seimbang [22]. Studi lain oleh Atlas et al. (2025) juga menerapkan klasifikasi dua kelas dan berhasil memperoleh performa model yang optimal dalam pengklasifikasian ulasan produk [2]. Ketiga, dari sisi teknis, penggunaan dua kelas memudahkan penerapan metode balancing data seperti oversampling dan resampling, serta mempercepat proses pelatihan model karena distribusi kelas lebih sederhana untuk diatur. Oleh karena itu, penggunaan dua kelas dianggap relevan dan efektif untuk mendukung tujuan dari penelitian ini dalam menilai sentimen ulasan pelanggan McDonald's di Australia secara lebih fokus dan terarah.

III. ANALISIS DAN RACANGAN

A. Preprocessing Data

Preprocessing data dilakukan sebelum analisis sentimen untuk memberikan hasil yang lebih akurat. Proses ini dilakukan dengan beberapa tahapan, yaitu:

- **Pembersihan Data:** Menghapus karakter khusus, tanda baca, dan angka yang tidak relevan.
- **Tokenisasi:** Memisahkan teks menjadi kata-kata individual agar dapat dianalisis lebih lanjut.
- **Stopword Removal:** Menghapus kata-kata umum yang tidak memberikan informasi penting.
- **Stemming dan Lemmatization:** Mengubah kata ke bentuk dasar agar variasi kata dapat dikelompokkan menjadi satu entitas yang sama.
- **Lowercasing:** Mengubah semua huruf menjadi huruf kecil supaya analisis dapat berjalan dengan lebih konsisten.

B. Analisis Sentimen

Analisis ini memiliki tujuan untuk mengeksplorasi sentimen pelanggan serta menentukan nilai dari ulasan pelanggan dengan nilai positif, negatif, atau netral. Hasil analisis sentimen dapat menentukan tingkat kepuasan pelanggan terhadap restoran. Hasil analisis sentiment dibagi menjadi tiga kategori utama, yaitu ulasan positif yang berisi apresiasi pelanggan kepada restoran, ulasan negatif yang berisi ketidakpuasan pelanggan dan keluhan pelanggan kepada restoran, serta ulasan netral yang tidak menunjukkan emosi atau apapun kepada restoran.

C. Evaluasi Model Sentimen

Evaluasi model sentimen dilakukan untuk mengukur akurasi dari model serta memberikan gambaran secara umum mengenai performa model dengan beberapa metode yang dapat digunakan seperti akurasi, *precision*, *recall*, *F1-score*, serta *confusion matrix*. Akurasi mengukur persentase prediksi yang benar dibandingkan total prediksi yang dilakukan, sehingga memberikan gambaran umum tentang performa model. *Precision* menghitung proporsi prediksi positif yang benar dibandingkan total prediksi positif yang dihasilkan model, yang penting dalam kasus di mana kesalahan klasifikasi positif lebih berdampak. *Recall* mengukur kemampuan model dalam mendeteksi ulasan dengan kategori sentimen yang benar, terutama dalam situasi di mana deteksi sentimen negatif atau positif sangat krusial. *F1-score* digunakan untuk menilai keseimbangan antara *precision* dan *recall*, sehingga dapat memberikan evaluasi terhadap performa model. *Confusion matrix* memberikan gambaran rinci tentang jumlah prediksi yang benar dan salah untuk setiap kategori sentimen, yang membantu dalam mengidentifikasi potensi bias dalam model. Dengan evaluasi ini, dapat diketahui metode mana yang memberikan hasil paling akurat dan dapat digunakan untuk menganalisis sentimen pelanggan.

D. Exploratory Data Analysis

Analisis sentimen ini digunakan untuk memahami distribusi data dan pola sentimen dalam ulasan pelanggan dengan beberapa tahapan. Tahapan pertama yang dilakukan adalah membuat word cloud yaitu mengidentifikasi kata-kata yang sering digunakan untuk menemukan topik atau aspek yang sering dibahas, dilanjutkan dengan mengidentifikasi kata-kata yang sering digunakan untuk menemukan ulasan yang dominan muncul dalam ulasan pelanggan. Model yang digunakan untuk menemukan topik utama dalam ulasan pelanggan adalah *Latent Dirichlet Allocation* (LDA), model ini dapat membantu menemukan topik utama yang sering dibahas oleh pelanggan serta mengelompokkan ulasan berdasarkan tema yang dibahas oleh pelanggan.

E. Visualisasi Data

Hasil analisis data divisualisasikan untuk mempermudah memahami hasil analisis sentimen, membantu memberikan hasil data secara lebih jelas, detail, dan intuitif. Dengan penyajian informasi yang lebih mudah dipahami, pengambilan keputusan dapat dilakukan dengan lebih efektif dan akurat. Visualisasi data ini dapat ditampilkan dalam bentuk word cloud, grafik dari distribusi sentiment, bar chart, serta histogram. Pustaka yang digunakan dalam visualisasi data mencakup Matplotlib, Seaborn, dan Word Cloud.

IV. IMPLEMENTASI

A. Persiapan Dataset

Tahap ini merupakan tahap pertama sebelum dilakukannya pemrosesan data, yaitu mempersiapkan *dataset*, membaca *dataset*, serta memuat pustaka yang akan digunakan sebelum dilakukannya analisis terhadap dataset. Pustaka yang digunakan diantaranya *Pandas*, *regex*, *Natural Language Toolkit* (NLTK), *TextBlob*, *Flair*, *Seaborn*, *Matplotlib*, *WordCloud*, *Scikit-learn*, serta *Transformers* dari *HuggingFace*. Dataset diimpor ke dalam *Python* menggunakan pustaka *pandas* dan disimpan pada sebuah *dataframe* bernama *df*.

- 1) Tampilan *Dataset*: Tahapan awal untuk mengenal *dataset* adalah dengan menampilkan beberapa baris pertama menggunakan fungsi *pandas* yaitu *df.head()*. Gambaran umum yang didapatkan dalam menampilkan *dataset* ini berbentuk struktur data, nama kolom, nilai awal data, serta format data.

	review_title	review_details	customer_name	review_timestamp
0	Horrible meal	Horrible mealHad meal at kelso nsw it was disg...	patricia	6h
1	Bad customer services, Rude staff	Bad customer services, Rude staffMc Donald's D...	jay	10h
2	Watch out for the "Karen", she's a government ...	Watch out for the "Karen", she's a government ...	Bazcar	17h
3	Morally Wrong	Morally WrongHow can Maccas find it morally ac...	FrankV	1d
4	McDonald's needs to pull its socks up - a lot	McDonald's needs to pull its socks up - a lotB...	Nika	1d

Gambar 1. Tampilan Data Teratas

- 2) Pada tahap ini, tipe data setiap kolom serta jumlah data yang tidak kosong diperiksa menggunakan fungsi `df.info()`. Analisis ini bertujuan untuk memeriksa dan mengidentifikasi tipe data setiap kolom, mengetahui jumlah baris yang memiliki nilai tidak kosong pada setiap kolom, serta mengetahui banyaknya data dengan nilai yang kosong.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 4 columns):
#   Column             Non-Null Count  Dtype
---  -
0   review_title        400 non-null    object
1   review_details      400 non-null    object
2   customer_name       400 non-null    object
3   review_timestamp    400 non-null    object
dtypes: object(4)
memory usage: 12.6+ KB
```

Gambar 2. Informasi Dataset

- 3) *Identifikasi Missing Value*: Untuk mendapatkan informasi yang lebih rinci mengenai data yang kosong (*missing value*) pada setiap kolom, fungsi `df.isnull().sum()` digunakan. Hasil analisis menunjukkan tidak terdapat nilai yang kosong dari setiap kolom.

```
review_title      0
review_details    0
customer_name     0
review_timestamp  0
```

Gambar 3. Hasil Identifikasi Missing Value

```
print("jumlah data duplikat :", df.duplicated().sum())
jumlah data duplikat : 1
```

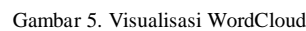
Gambar 4. Jumlah Data Duplikat Pada Dataset

- 4) *Deteksi Duplikasi Data*: Pada tahap ini, dilakukan pemeriksaan jumlah data duplikat dalam *dataset* menggunakan fungsi `df.duplicated().sum()`. Berdasarkan hasil analisis, terdapat satu baris data duplikat. Namun, data tersebut tidak dihapus karena jumlahnya sangat sedikit dan tidak dianggap memberikan pengaruh signifikan terhadap hasil analisis secara keseluruhan.

B. Pra-pemrosesan Teks

Tahap pra-pemrosesan teks bertujuan untuk menyiapkan data ulasan dalam bentuk yang dapat diproses oleh model analisis dengan lebih efektif. Proses ini melibatkan pembersihan teks dari elemen-elemen yang tidak relevan, normalisasi format teks, serta konversi data ke dalam bentuk numerik. Kata-kata umum yang tidak memiliki nilai informasi penting, atau disebut *stop words*, diidentifikasi dan dihapus menggunakan daftar dari pustaka NLTK. Selain itu, seluruh teks diubah menjadi huruf kecil untuk menghindari perbedaan makna antara kata-kata yang serupa seperti “Food” dan “food”. Aksentuasi atau karakter khusus dalam teks juga dihapus dan disederhanakan ke dalam bentuk ASCII melalui parameter `strip_accents='ascii'`. Setelah tahap pembersihan selesai, data teks kemudian ditransformasikan menjadi representasi numerik menggunakan metode TF-IDF (*Term Frequency–Inverse Document Frequency*) melalui fungsi `TfidfVectorizer` dari pustaka *Scikit-learn*. Metode ini memberikan bobot pada setiap kata berdasarkan seberapa sering kata tersebut muncul dalam satu dokumen dibandingkan dengan seluruh korpus, dengan mengaktifkan parameter `use_idf=True`. Parameter `stop_words` juga diatur untuk mengecualikan kata-kata yang telah didefinisikan sebelumnya. Hasil dari tahapan ini adalah vektor numerik yang merepresentasikan setiap dokumen teks, yang kemudian dapat digunakan sebagai input dalam proses pelatihan dan evaluasi model klasifikasi sentimen.

Setelah proses pembersihan dan transformasi teks menggunakan TF-IDF, dilakukan visualisasi awal menggunakan WordCloud untuk mengidentifikasi kata-kata yang paling sering muncul dalam ulasan pelanggan. Kolom `review_title` dan `review_details` digabung agar konteks ulasan lebih utuh, lalu dibersihkan dari *stop words* dan disatukan menjadi satu teks panjang. Visualisasi dibuat menggunakan pustaka WordCloud dengan latar belakang putih dan menghapus kata-kata umum agar fokus pada kata bermakna informatif. WordCloud ini menggunakan pendekatan *unigram* untuk menampilkan kata-kata tunggal yang dominan secara visual. Hasilnya menunjukkan bahwa kata seperti *time*, *store*, *staff*, dan *service* paling sering muncul, mencerminkan perhatian pelanggan terhadap pelayanan, pemesanan, dan kualitas makanan. Visualisasi ini memberikan gambaran awal yang berguna sebelum dilakukan analisis sentimen lebih lanjut.



D. Analisis Sentimen Berdasarkan Voting Antar Model

Setelah seluruh model dianalisis secara individu, dilakukan pendekatan voting untuk menentukan label sentimen akhir berdasarkan hasil keempat model. Langkah awalnya adalah mengevaluasi konsistensi prediksi antar model dengan membuat kolom khusus yang menandai apakah keempat model menghasilkan label yang sama. Selain itu, diterapkan metode voting sederhana untuk menghasilkan label sentimen final. Jika terdapat mayoritas (minimal tiga model) yang sepakat, maka label tersebut digunakan sebagai hasil akhir. Namun, apabila terjadi hasil imbang antara label 'POSITIVE' dan 'NEGATIVE', maka ulasan tersebut dikategorikan sebagai 'DRAW', yang menandakan ketidaksepakatan antar model.

Untuk menangani kasus 'DRAW' ini, dilakukan peninjauan manual terhadap ulasan yang masuk kategori tersebut. *Dataset* dengan label 'DRAW' disaring dan disimpan ke dalam *file* eksternal agar dapat ditinjau dan diberi label ulang secara manual. Setelah proses *labeling* manual selesai, hasil tersebut diintegrasikan kembali ke dalam *dataset* utama dengan mengganti label 'DRAW' menggunakan hasil *labeling* terbaru. Proses ini memastikan bahwa seluruh *data* memiliki label sentimen yang pasti dan konsisten, sehingga *dataset* akhir siap digunakan untuk tahap analisis atau pelatihan model selanjutnya tanpa adanya nilai ambigu.

Sebagai hasil dari proses voting antar model dan pelabelan ulang data yang ambigu, diperoleh *dataset* akhir yang lebih bersih dan konsisten untuk digunakan dalam pelatihan model klasifikasi. Pendekatan voting tidak hanya meningkatkan keandalan label sentimen, tetapi juga merepresentasikan hasil kolektif dari beberapa pendekatan analisis sentimen yang berbeda, sehingga menghasilkan anotasi data yang lebih akurat. *Dataset* ini kemudian digunakan untuk membangun dan menguji performa dua algoritma klasifikasi, yaitu *Naive Bayes Multinomial* dan *Support Vector Machine (SVM)*, guna mengevaluasi kemampuan model dalam mengklasifikasikan sentimen secara otomatis. Tahapan ini menjadi jembatan penting antara proses analisis awal dan pembuatan sistem klasifikasi yang andal berbasis pembelajaran mesin.

E. Penanganan Duplikasi dan Ketidakseimbangan Label

Sebelum pelatihan model dilakukan, langkah awal yang diambil adalah membersihkan *dataset* dari duplikasi, khususnya pada kolom *review_details*, guna mencegah bias dan *overfitting*. Mengingat distribusi data yang tidak seimbang, di mana kelas negatif jauh lebih dominan, dilakukan dua skenario pelatihan dengan teknik *oversampling*. Skenario pertama menggunakan seluruh data latih tanpa pemisahan data uji, namun menghasilkan akurasi sangat tinggi yang mengindikasikan *overfitting*. Untuk mengatasi hal tersebut, skenario kedua memisahkan sebagian data positif sebagai data uji dan menerapkan *oversampling* hanya pada data latih, sehingga evaluasi model menjadi lebih realistis.

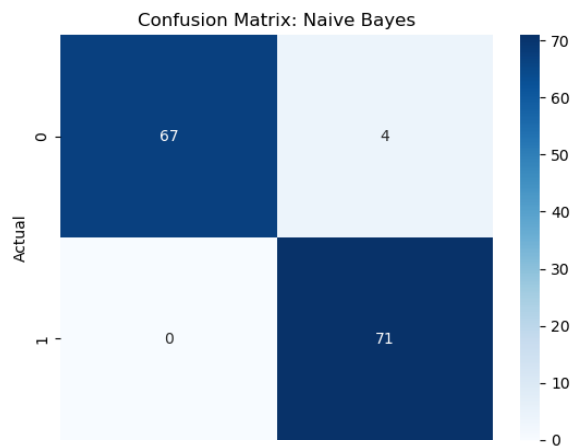
F. Oversampling dan Pelatihan Model

Pada tahap evaluasi model, *Multinomial Naive Bayes* menunjukkan performa yang stabil dan cocok untuk data berbasis frekuensi kata seperti hasil TF-IDF, namun cenderung kurang fleksibel terhadap pola kompleks. Sementara itu, model SVC dengan kernel linear menunjukkan akurasi yang sangat tinggi, bahkan mencapai 100% dalam beberapa percobaan awal, yang mengindikasikan potensi *overfitting*. Untuk mengonfirmasi hal ini, diterapkan validasi silang lima fold yang menghasilkan akurasi tinggi dengan standar deviasi sangat rendah, menandakan bahwa model sangat konsisten namun perlu diuji lebih lanjut pada data yang lebih beragam. Oleh karena itu, analisis tidak hanya berfokus pada akurasi semata, tetapi juga memperhatikan kestabilan performa model antar fold serta kemungkinan bias akibat karakteristik *dataset* yang homogen.

- 1) *Hasil Pelatihan Naive Bayes*: Model *Multinomial Naive Bayes* yang dilatih menggunakan data hasil *oversampling* menunjukkan performa klasifikasi yang sangat baik. Berdasarkan hasil evaluasi pada data uji, model ini menghasilkan akurasi sebesar 95.8%, *precision* sebesar 94.7%, *recall* sebesar 97.2%, dan *F1-score* sebesar 95.9%. *Confusion matrix* memperlihatkan bahwa dari total 142 data uji, hanya 4 data yang diklasifikasikan secara salah, semuanya berasal dari kelas negatif yang salah diklasifikasikan sebagai positif. Hasil ini menunjukkan bahwa model cukup sensitif terhadap kelas minoritas, namun tetap ada risiko *overfitting* mengingat proses pelatihan dilakukan setelah *oversampling*.

```
=== Naive Bayes ===  
Accuracy: 0.971830985915493  
Precision: 0.9466666666666667  
Recall: 1.0  
F1-score: 0.9726027397260273
```

Gambar 8. Hasil Perhitungan Metrik Naive Bayes



Gambar 10. Visualisasi Confusion Matrix Naive Bayes

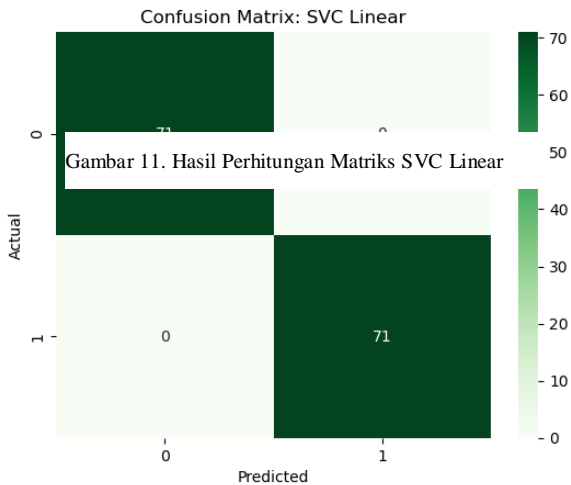
2) *Hasil Cross-Validation Naive Bayes:* Untuk menguji kestabilan model, dilakukan validasi silang sebanyak lima *fold*. Hasilnya menunjukkan rata-rata akurasi sebesar 95.63%, dengan nilai per lipatan sebesar [94.37%, 97.18%, 91.55%, 97.16%, 97.87%]. Standar deviasi sebesar 2.37% menandakan adanya sedikit variasi performa antar subset data. Interval kepercayaan 95% terhadap akurasi berada di kisaran 93.55% hingga 97.70%. Ini mengindikasikan bahwa model relatif stabil dan handal, meskipun tetap perlu diuji pada data yang benar-benar baru.

```
=== Naive Bayes Cross-Validation ===
Mean Accuracy: 0.9563
All fold scores: [0.94366197 0.97183099 0.91549296 0.97163121 0.9787234 ]
Standard Deviation (Naive Bayes): 0.0237
Confidence Interval (Naive Bayes, 95%): (0.9355110007950509, 0.9770252093737635)
```

Gambar 9. Hasil Perhitungan Cross-Validation Naive Bayes

3) *Hasil Pelatihan SVC Linear:* Model SVC Linear memberikan hasil yang sangat tinggi dan konsisten. Evaluasi pada data uji menghasilkan akurasi, *precision*, *recall*, dan *F1-score* yang semuanya bernilai sempurna (1.0). *Confusion matrix* menunjukkan bahwa seluruh data uji berhasil diklasifikasikan secara benar tanpa kesalahan satu pun. Hasil ini sangat impresif namun juga mengundang perhatian terhadap kemungkinan *overfitting*, terutama karena data yang digunakan telah mengalami proses *oversampling* secara penuh.

```
=== SVC Linear ===
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1-score: 1.0
```



Gambar 11 Visualisasi Confusion Matriks SVC Linear

- 4) *Hasil Cross-Validation SVC Linear*: Pada evaluasi menggunakan validasi silang 5-fold, model *SVC Linear* juga menunjukkan hasil sempurna dengan rata-rata akurasi 100% di seluruh lipatan ([1.0, 1.0, 1.0, 1.0, 1.0]) dan standar deviasi nol. Namun, tidak dapat dihitungnya *confidence interval* (karena deviasi nol) mengindikasikan bahwa performa model terlalu homogen, dan besar kemungkinan *overfitting* terjadi. Ini memperkuat kebutuhan untuk menguji model terhadap data uji eksternal yang tidak terpengaruh oleh proses *oversampling* atau pelatihan untuk menilai kemampuan generalisasi secara realistis.

```
=== SVC Linear Cross-Validation ===
Mean Accuracy: 1.0000
All fold scores: [1. 1. 1. 1. 1.]
Standard Deviation (SVC): 0.0000
```

Gambar 12 Hasil Perhitungan Cross-Validation SVC Linear

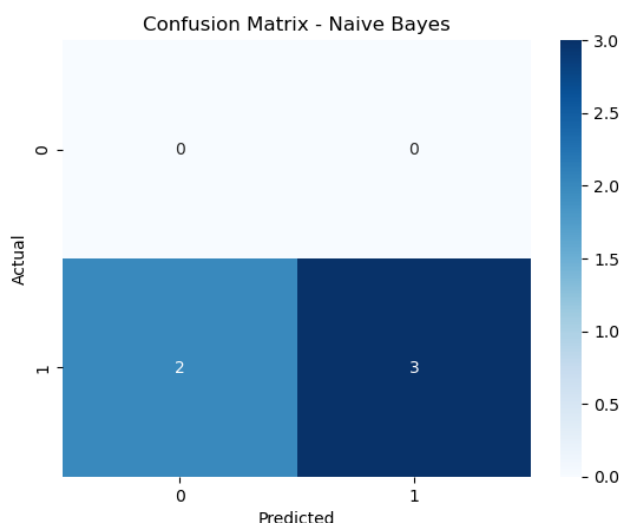
G. Pelatihan Model dengan Pemisahan Sebagian data

Pendekatan pelatihan kedua dirancang untuk menghindari *overfitting* dengan memisahkan sebagian data positif sebagai data uji yang tidak dilibatkan dalam proses pelatihan. Sebanyak lima data positif dipilih secara acak dan disisihkan sebagai *test set*, sementara sisa data positif digabungkan dengan seluruh data negatif membentuk *training set*. Setelah dilakukan transformasi teks menjadi fitur numerik dengan *TF-IDF*, label dikodekan secara biner agar dapat digunakan dalam model klasifikasi. Untuk menangani ketidakseimbangan data pada *training set*, diterapkan teknik *Random OverSampling* menggunakan pustaka *imbalanced-learn* untuk menyamakan jumlah data antara kelas mayoritas dan minoritas sebelum dilakukan pelatihan model.

- 1) *Hasil Evaluasi Model Naive Bayes*: Model *Multinomial Naive Bayes* dilatih menggunakan data latih hasil *oversampling* dan diuji menggunakan lima data positif yang terpisah. Hasil evaluasi menunjukkan akurasi sebesar 60%, *precision* 60%, *recall* sempurna 100%, dan *F1-score* 75%. Ini mengindikasikan bahwa model berhasil mengenali seluruh data positif (tanpa false negative), namun *precision* yang rendah menunjukkan adanya prediksi positif yang salah. Hal ini terlihat pada confusion matrix, di mana 3 data diprediksi benar sebagai positif, sementara 2 data lainnya salah diklasifikasikan sebagai negatif. Karena jumlah data uji hanya lima dan seluruhnya berasal dari kelas positif, hasil ini memberikan gambaran awal tentang potensi generalisasi model, meski belum cukup representatif.

```
=== Naive Bayes ===
Accuracy: 0.6
Precision: 1.0
Recall: 0.6
F1-score: 0.7499999999999999
```

Gambar 13. Hasil Perhitungan Metrik Naive Bayes dengan 5 Data Terpisah



Gambar 14. Visualisasi Confusion Matrix dengan 5 Data Terpisah

- 2) *Hasil Evaluasi Model SVC Linear*: Model SVC dengan *kernel linear* juga diuji menggunakan strategi yang sama. Hasil evaluasi menunjukkan akurasi 60%, *precision* sempurna 100%, *recall* 60%, dan *F1-score* 75%. Berbeda dengan *Naive Bayes*, model SVC tidak menghasilkan prediksi positif yang salah (*false positive*), namun gagal mengenali dua dari lima data positif (*false negative*). Hal ini tercermin dari *confusion matrix* yang memperlihatkan 3 prediksi benar

```
Accuracy: 0.6
Precision: 1.0
Recall: 0.6
F1-score: 0.7499999999999999
```

Confusion Matrix - SVC

	Predicted 0	Predicted 1
Actual 0	0	0
Actual 1	2	3

The heatmap visualizes the confusion matrix for an SVC model. The color scale ranges from 0.0 (light yellow) to 3.0 (dark green). The matrix shows 0 correct classifications for Actual 0, 2 misclassifications for Actual 1 (Predicted 0), and 3 correct classifications for Actual 1 (Predicted 1).

[illegible]

69

-

Topik 1:
big | mcdonald | salad | burgers | ordered | mac | would | order | cold | back

Topik 2:
coffee | maccas | good | need | shake | basket | bathroom | busy | visit | cookie

Topik 3:
time | get | thanks | staff | service | amazinghello | sample | free | food | making

Topik 4:
bk | salty | sugar | maccas | yuck | incredibly | superiorbig | service | outstanding | far

Topik 5:
good | team | store | came | deserves | grimce | cenat | kai | toilet | 10

70

DAFTAR PUSTAKA

- [1] N. Aula, M. Ula dan L. Rosnita, "ANALISIS SENTIMEN REVIEW CUSTOMER TERHADAP PERUSAHAAN," *Journal of Informatics and Computer Science*, vol. 9, no. 1, pp. 82-83, 2023.
- [2] L. G. Atlas, D. Arockiam, A. Muthusamy, B. Balusamy, S. Selvarajan, T. Al-Shehari dan N. A. Alsadhan, "A modernized approach to sentiment analysis of product reviews using BiGRU and RNN based LSTM deep learning models," *Scientific Reports*, vol. 15, no. 1, p. 16642, 2025.
- [3] M. Kayed, R. P. Díaz-Redondo dan A. Mabrouk, "Deep Learning-based Sentiment Classification: A Comparative Survey," *arXiv preprint arXiv:2312.17253*, 2023.
- [4] Q. Kong, T. Siauw dan A. M. Bayen, "Why Python?," dalam *Python Programming and Numerical Methods: A Guide for Engineers and Scientists*, Boston, Academic Press, 2020, p. xxii.
- [5] A. K. Putri dan D. I. Nur, "PENGUNAAN BAHASA PYTHON UNTUK ANALISIS DAN VISUALISASI DATA PENDUDUK DI DESA SUMBERJO, NGANJUK," *Jurnal Pengabdian Kepada Masyarakat*, vol. 3, no. 3, pp. 207-208, 2023.
- [6] M. Radhi, A. D. R. H. Sitompul, S. H. Sinurat dan E. Indra, "ANALISIS BIG DATA DENGAN METODE EXPLORATORY DATA ANALYSIS(EDA) DAN METODE VISUALISASI MENGGUNAKAN JUPYTER NOTEBOOK," *JUSIKOM PRIMA (Jurnal Sistem InformasidanIlmu Komputer Prima)*, vol. 4, no. 2, p. 24, 2021.
- [7] W. McKinney, *Data Analysis with Pandas and Python*, Sebastopol: O'Reilly Media, 2022.
- [8] J. Reback, *Pandas: Powerful Python Data Analysis Toolkit*, Austin: Pandas Development Team, 2023.
- [9] Y. Chen dan L. Zhang, "Advanced Data Processing Using Pandas," *Journal of Data Science and Machine Learning*, vol. 12, no. 1, pp. 55-72, 2024.
- [10] M. Jojoa, P. Eftekhar, B. Nowrouzi Kia dan B. Garcia Zapiain, "Natural language processing analysis applied to COVID 19 open text opinions using a distilBERT model for sentiment categorization," *AI & SOCIETY*, vol. 39, no. 3, pp. 883-890, 2024.
- [11] D. Anggraini, S. Rahmawati dan R. Kurniawan, "Natural Language Processing For Automatic Sentiment Analysis In Social Media Data," *International Journal of Information Engineering and Science*, vol. 1, no. 1, pp. 16-19, 2024.
- [12] M. S. Başarslan dan F. Kayaalp, "Sentiment Analysis with Machine Learning Methods on Social Media," *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 9, no. 3, pp. 5-15, 2020.
- [13] J. Setiawan, V. Gousander dan I. Prasetyawan, "Unmasking the Sentiments of Labuan Bajo: An Instagram-based Analysis for Tourism Insights through VADER Sentiment Analysis," *G-Tech : Jurnal Teknologi Terapan*, vol. 7, no. 3, pp. 967-976, 2023.
- [14] M. A. Gumilang, F. Abdillah, M. Y. Amin dan M. Hasan, "Analisis Sentimen Media Sosial Kementerian RI: Respons Warganet Memanfaatkan TextBlob Analyzer," *Jurnal Sosioteknologi*, vol. 23, no. 2, pp. 203-213, 2024.
- [15] M. Javed, S. A. Parveen dan S. K. Singh, "Hybrid deep learning model for sentiment analysis," *Procedia Computer Science*, vol. 185, pp. 377-384, 2021.
- [16] I. H. Sarker, "Sentiment analysis of social media posts using deep learning models," *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 12743-12765, 2021.
- [17] M. Gaur, A. Sheth, N. Doshi, C. Thomas, V. Chan dan P. Kapanipathi, "Sentiment Analysis and Visualization of COVID-19 Tweets using NLP Techniques," *2021 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1-5, 2021.
- [18] M. L. Waskom, "Seaborn: Statistical Data Visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.
- [19] S. Chatterjee, "Enhancing Data Visualization with Seaborn," *Data Science Journal*, vol. 15, no. 2, pp. 45-58, 2024.
- [20] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Computing in Science & Engineering*, vol. 25, no. 1, pp. 90-95, 2023.
- [21] P. Barrett, J. D. Hunter dan J. M. T., "Matplotlib for Data Visualization," *Scientific Computing Journal*, vol. 18, no. 3, pp. 23-40, 2025.
- [22] I. S. S. Shanto, Z. Ahmed, N. Hossain, A. Roy dan A. I. Jony, "Binary vs. Multiclass Sentiment Classification for Bangla E-commerce Product Reviews: A Comparative Analysis of Machine Learning Models," *International Journal of Information Engineering and Electronic Business (IJIEEB)*, vol. 15, no. 6, p. 48-63, 2023.