PENERAPAN *SPARK* DENGAN METODE *NAIVE BAYES CLASSIFIER* PADA ANALISIS SENTIMEN PENGUNJUNG KE PANTAI RIO BY THE BEACH DI *GOOGLE MAPS* KANTOR PERWAKILAN BANK INDONESIA PROVINSI LAMPUNG

Disusun oleh:
ELSYAH SAPYRAH
121450096

PROGRAM STUDI SAINS DATA

FAKULTAS SAINS

INSTITUT TEKNOLOGI SUMATERA

2024

Elsyah Sapyrah

Sains Data

Penerapan *Spark* dengan Metode *Naive Bayes Classifier* Pada Analisis Sentimen Pengunjung ke Pantai Rio By The Beach di *Google Maps*

ABSTRAK

Sektor pariwisata merupakan pendorong pertumbuhan ekonomi daerah. Di Provinsi Lampung, upaya peningkatan ekonomi dilakukan melalui pengembangan destinasi wisata seperti Pantai Rio By The Beach. Penelitian ini bertujuan untuk menganalisis tingkat kepuasan pengunjung terhadap pantai tersebut melalui ulasan di Google Maps dengan menggunakan metode Naive Bayes. Data rating dan ulasan dikumpulkan dan dianalisis menggunakan Spark dengan berbagai tahapan, yaitu pengumpulan data, praproses data, pelatihan model, dan evaluasi hasil. Hasil menunjukkan bahwa metode Naive Bayes mencapai akurasi 83% dalam mengklasifikasikan ulasan menjadi positif dan negatif. Dengan K-fold Cross Validation (k=7), akurasi meningkat menjadi 89%. Dari 350 ulasan, diperoleh hasil sentimen 90% positif dan 10% negatif. Faktor positif utama meliputi keindahan pantai, kebersihan, pasir putih, harga tiket, dan fasilitas. Faktor negatif mencakup akses jalan, parkir, dan tempat berteduh. Penelitian ini memberikan wawasan untuk meningkatkan kualitas layanan dan daya tarik Pantai Rio By The Beach serta destinasi wisata lainnya di Lampung.

Kata Kunci: ***Analisis Sentimen, Naive Bayes, Spark, Pariwisata, Pantai Rio By The Beac***

DAFTAR ISI

LEMBAR PENGESAHAN.....i
ABSTRAK.....ii
KATA PENGANTAR.....iii
DAFTAR ISI.....iv
DAFTAR TABEL.....v
DAFTAR GAMBAR.....vi
DAFTAR LAMPIRAN.....vii
BAB I PENDAHULUAN.....1
A. Latar Belakang.....*1*
B. Rumusan Masalah.....*2*
C. Tujuan.....*2*
BAB II TINJAUAN PUSTAKA.....3
A. Studi Terdahulu.....*3*
B. Definisi Spark.....*3*
C. Definisi Naive Bayes Classifier.....*4*
D. Definisi Analisis Sentimen.....*4*
BAB III METODOLOGI.....6
A. Waktu dan Tempat Pelaksanaan Kerja Praktik.....*6*
B. Diagram Alir.....*6*
C. Metode Pengolahan Data.....*6*
a. Pengambilan Data.....*6*
b. Preprocessing Data.....*7*
i. Data Cleaning.....*7*
ii. Manipulation Data.....*7*
iii. Labelling.....*7*
iv. Case Folding.....*8*
v. Tokenizer.....*8*
vi. StopWord Removal.....*8*
c. Ekstraksi Fitur.....*8*
i. Count Vectorizer.....*8*
ii. TF-IDF.....*9*
d. Splitting Data.....*9*
e. Modelling.....*9*
i. Naive Bayes.....*9*

ii. Visualisasi Word Cloud..... 10
f. Evaluasi Model..... 10
BAB IV HASIL DAN PEMBAHASAN.....12
A. Deskripsi Data.....12
B. Labelling.....12
C. Model Naive Bayes.....13
D. Evaluasi Model K-fold Cross Validation.....13
E. Visualisasi.....14
a. Word Cloud Positif.....14
b. Word Cloud Negatif..... 14
BAB V PENUTUP.....16
A. Kesimpulan.....16
B. Saran.....16
DAFTAR PUSTAKA.....17

DAFTAR TABEL

Tabel 1. Penerapan *Manipulation Data*.....7
Tabel 2. Penerapan *Labelling*.....8
Tabel 3. Penerapan *Case Folding*.....8
Tabel 4. Penerapan *Tokenizer*.....8
Tabel 5. Penerapan *Count Vectorizer*.....9
Tabel 6. Penerapan TF-IDF.....9
Tabel 7. *Confusion Matrix*.....10
Tabel 8. Variabel pada data *review* pengunjung Pantai Rio By The Beach.....12
Tabel 9. Jumlah Sentimen Pengunjung Pantai Rio By The Beach.....13
Tabel 10. *Confusion Matrix* Model *Naive Bayes*.....13
Tabel 11. Hasil Evaluasi Model *Naive Bayes*.....13
Tabel 12. *Confusion Matrix* Model *K-fold Cross Validation* dengan k=7……13
Tabel 13. Hasil Evaluasi Model *K-fold Cross Validation* dengan k=7……13

DAFTAR GAMBAR

Gambar 1. Diagram Alir.....6
Gambar 2. Dataset.....7
Gambar 3. *Word Cloud* Positif.....14
Gambar 4. *Word Cloud* Negatif.....15

DAFTAR LAMPIRAN

Lampiran 1. Form Pengajuan Surat Pengantar Kerja Praktik.....18
Lampiran 2. Permohonan Surat Pengantar Magang.....19
Lampiran 3. Laporan Kerja Harian.....20
Lampiran 4. Laporan Bimbingan.....26

BAB I PENDAHULUAN

A. Latar Belakang

Sektor pariwisata merupakan salah satu sektor yang mendorong laju pertumbuhan ekonomi di suatu daerah [1]. Dalam upaya meningkatkan pertumbuhan ekonomi di Provinsi Lampung, Pemerintah Daerah bekerjasama dengan Kantor Perwakilan Bank Indonesia Provinsi Lampung membentuk Forum Investasi Lampung (FOILA) dalam rangka menarik minat investor untuk berinvestasi di berbagai sektor di Provinsi Lampung, termasuk pariwisata [2]. Dengan adanya infrastruktur yang mendukung, maka dapat mewujudkan peran sektor pariwisata dalam meningkatkan pertumbuhan ekonomi di Provinsi Lampung.

Provinsi Lampung sendiri memiliki beragam kekayaan alam yang indah dan sering dikunjungi masyarakat baik dari dalam maupun luar daerah, salah satunya adanya wisata pantai. Destinasi wisata yang tengah populer di Lampung adalah Pantai Rio By The Beach, sebuah wisata pantai baru yang terletak di Lampung Selatan. Pantai ini berhasil menarik perhatian banyak pengunjung karena keindahan alamnya dan fasilitas yang semakin lengkap dengan harga tiket masuk yang memadai. Sebagai destinasi wisata yang berkembang, Pantai Rio By The Beach memiliki potensi besar untuk memberikan kontribusi positif terhadap perekonomian lokal terutama bagi masyarakat setempat yang dapat memanfaatkan peluang ekonomi dari sektor pariwisata.

Untuk meningkatkan kualitas layanan dan daya tarik pengunjung diperlukan upaya kerjasama Pemerintah Daerah dan pelaku industri pariwisata. Salah satu langkah penting adalah memahami tingkat kepuasan pengunjung terhadap infrastruktur Pantai Rio By The Beach melalui *rating* dan *review* di *Google Maps*. Analisis sentimen adalah proses dalam menganalisis sebuah teks digital untuk menentukan apakah nada emosional dari ulasan tersebut positif, negatif, atau netral [3].

Salah satu metode yang populer untuk melakukan analisis sentimen adalah *Naive Bayes* yang merupakan salah satu metode dalam *supervised learning* yang biasanya digunakan untuk klasifikasi. Naive Bayes adalah algoritma analisis statistik yang mengolah data numerik dengan probabilitas Bayesian agar dapat memprediksi suatu kelas anggota. Naive Bayes mengasumsikan bahwa suatu nilai atribut sebuah kelas tidak dipengaruhi atau mempengaruhi suatu nilai atribut lainnya [4].

Dalam penelitian ini, data *rating* dan *review* pengunjung ke Pantai Rio By The Beach dari *Google Maps* akan dikumpulkan dan dianalisis menggunakan metode Naive Bayes. Proses analisis akan mencakup pengumpulan data, praproses data, pelatihan model, dan evaluasi hasil analisis. Ulasan yang diberikan oleh pengunjung di platform seperti *Google Maps* dapat menjadi indikator penting untuk memahami persepsi dan pengalaman mereka. Melalui analisis ulasan ini diharapkan dapat memberikan informasi berharga mengenai aspek-aspek yang perlu ditingkatkan serta kekuatan yang harus dipertahankan Pantai Rio By The Beach, serta destinasi wisata lainnya di Provinsi Lampung.

B. Rumusan Masalah

1. Bagaimana hasil dari penerapan metode *Naive Bayes Classifier* untuk mengklasifikasikan sentimen ulasan pengunjung terhadap Pantai Rio By The Beach?
2. Bagaimana tingkat kepuasan pengunjung terhadap Pantai Rio By The Beach di Provinsi Lampung berdasarkan ulasan yang diberikan di *Google Maps*?
3. Apa saja faktor-faktor utama yang mempengaruhi sentimen positif dan negatif dari ulasan pengunjung terhadap Pantai Rio By The Beach?

C. Tujuan

1. Mengetahui hasil dari penerapan *Spark* dengan Metode *Naive Bayes Classifier* untuk klasifikasi sentimen ulasan pengunjung terhadap Pantai Rio By The Beach di *Google Maps*
2. Menentukan tingkat kepuasan pengunjung terhadap Pantai Rio By The Beach
3. Menganalisis faktor-faktor utama yang mempengaruhi sentimen positif dan negatif dari ulasan pengunjung terhadap Pantai Rio By The Beach

BAB II TINJAUAN PUSTAKA

A. Studi Terdahulu

Ada beberapa studi terdahulu yang relevan dengan analisis sentimen *review* pengunjung wisata melalui *google maps* menggunakan metode Naive Bayes. Beberapa hasil penelitian tersebut diperoleh dari basis data jurnal online karena telah dipublikasikan sebelumnya, yaitu :

1. Studi yang dilakukan oleh Gergorius Kopong Pati, Elfira Umar yang berjudul "Analisis Sentimen Komentar Pengunjung Terhadap Tempat Wisata Danau Weekuri Menggunakan Metode *Naive Bayes Classifier* Dan *K Nearest Neighbor*" yang dipublikasikan pada Jurnal Media Informatika Budidarma Vol. 6, Nomor 4, Tahun 2022. Penelitian ini berisi mengenai perbandingan metode *Naive Bayes* dan *K Nearest Neighbor* dalam menganalisis sentimen pengunjung ke wisata Danau Weekuri. Dari pengujian akurasi menggunakan metode *K-Nearest Neighbor* dimana diperoleh tingkat akurasinya 76.53%, sedangkan tingkat akurasi dengan metode *Naive Bayes Classifier* Sebesar 73.47%. Oleh karena itu tingkat akurasi yang diperoleh dengan menggunakan metode KNN lebih baik
2. Studi yang dilakukan oleh Nabila Aurelia Rahma, Garno, Nina Sulistiyowati yang berjudul "Analisis Sentimen Tempat Wisata Di Jakarta Pasca Covid -19 Dengan Algoritma *Naïve Bayes*" yang dipublikasikan pada Jurnal Pendidikan dan Konseling Vol. 4, Nomor 6, Tahun 2022. Penelitian ini berisi mengenai analisis dari metode *Naive Bayes* pada sentimen pengunjung ke wisata di Jakarta pasca Covid-19. Diperoleh kesimpulan bahwa penerapan algoritma *Naive Bayes* memberikan hasil yaitu 1679 label positif, 534 label negatif, dan 1211 label netral. Sentimen pengunjung di media sosial twitter lebih banyak sentimen positif dibanding sentimen negatif., sentimen netral walaupun mendominasi dari total data tidak berpengaruh terhadap peningkatan kualitas tempat pariwisata di Jakarta.
3. Studi yang dilakukan oleh Julius Chrisostomus Aponno yang berjudul "Penerapan Algoritma *Sentiment Analysis* Dan *Naïve Bayes* Terhadap Opini Pengunjung Di Tempat Wisata Pantai Pintu Kota, Kota Ambon" yang dipublikasikan pada Jurnal Teknik Informatika dan Sistem Informasi Vol. 9, Nomor 6, Tahun 2022. Penelitian ini berisi mengenai analisis dari metode Naive Bayes pada sentimen pengunjung ke wisata di Pantai Pintu Kota, Kota Ambon. Dari penelitian tersebut dengan metode

Naive Bayes diperoleh akurasi 90.65% Ini menunjukkan bahwa nilai akurasi dengan menggunakan metode Naive Bayes baik ini dikarenakan nilai TP dengan TN memiliki nilai tidak terlampau jauh dengan nilai TP 83.33% dan TN 96.43%

B. Definisi *Spark*

Apache Spark adalah *platform* komputasi klaster sumber terbuka yang dirancang untuk komputasi data yang cepat dan serbaguna. *Apache Spark* diketahui sebagai *tools big* data yang mempunyai kecepatan 1000 kali lebih cepat daripada *hadoop* dan *Mlib. Spark* adalah mesin pembelajaran dari *Apache Spark* dengan *library* algoritma yang cukup lengkap untuk proses *analytic* [5]. Pada penelitian ini digunakan bahasa pemrograman Python.

C. Definisi *Naive Bayes Classifier*

Naive Bayes Classification adalah teknik klasifikasi yang didasarkan pada Teorema Bayes dengan asumsi adanya independensi di antara prediktor. *Naive Bayes Classifier* memprediksi probabilitas masa depan berdasarkan pengalaman sebelumnya yang dikenal juga sebagai Teorema Bayes. Secara sederhana, klasifikasi Naive Bayes mengasumsikan bahwa keberadaan suatu fitur tertentu dalam suatu kelas tidak bergantung pada keberadaan fitur lainnya.

Keuntungan dari penggunaan metode ini adalah bahwa ia hanya memerlukan sejumlah kecil data pelatihan untuk menentukan estimasi parameter yang diperlukan dalam proses klasifikasi. Karena metode ini mengasumsikan variabel-variabel bersifat independen, hanya varians dari suatu variabel dalam sebuah kelas yang diperlukan untuk menentukan klasifikasi, bukan keseluruhan matriks kovarians [6].

Perhitungan algoritma Naive Bayes pada Persamaan 1 dan perhitungan prior direpresentasikan pada persamaan 2.

$$P(d) = P(c) * P(d|c) \quad (1)$$

$$P(c) = \frac{N_c}{N} \quad (2)$$

Keterangan:

$P(c|d)$: Posterior atau Probabilitas kelas c diberikan dokumen d

$P(c)$: Prior atau Probabilitas awal muncul kategori c

$P(d|c)$: Likelihood

N_c : Jumlah dokumen kelas c

N : Jumlah seluruh dokumen

D. Definisi Analisis Sentimen

Analisis sentimen adalah riset komputasional terhadap opini, sentimen, dan emosi yang diekspresikan dalam teks, kemudian diklasifikasikan menjadi kelompok sentimen positif dan negatif. Dalam penelitian ini, saya menggunakan kategori sentimen *Fined* untuk mengklasifikasikan data opini menjadi kalimat positif atau negatif.

Secara umum, analisis sentimen dibagi menjadi dua kategori utama, yaitu:

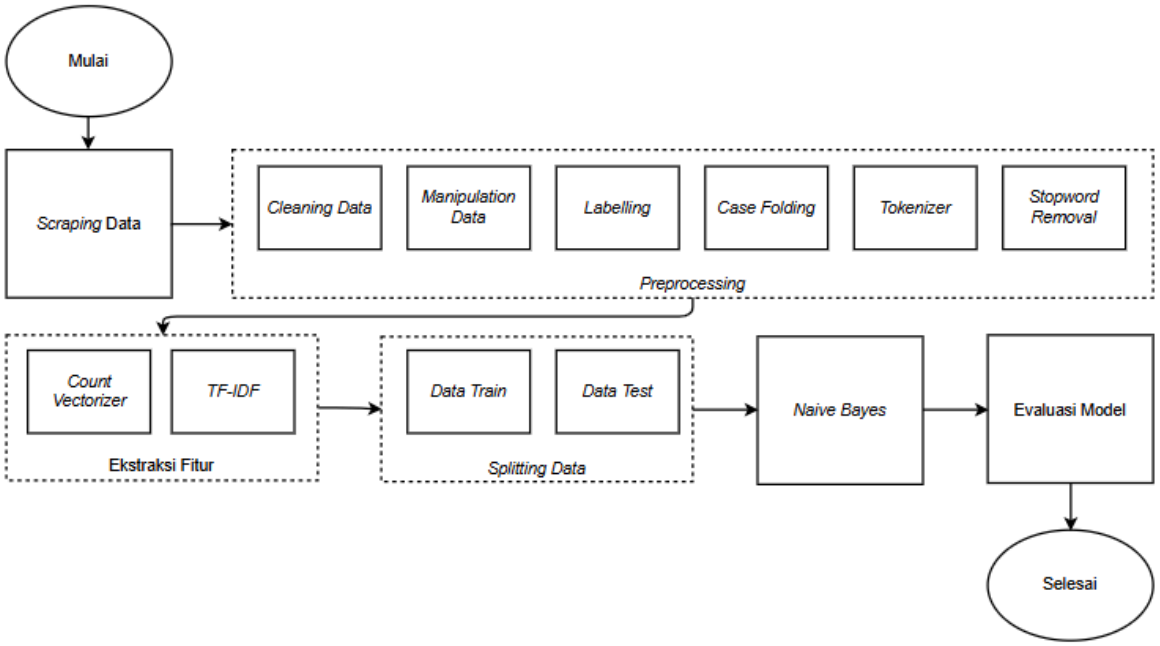
1. *Coarse - grained sentiment analysis*: Proses analisis dan klasifikasi orientasi sebuah dokumen secara keseluruhan. Orientasi ini dapat dibagi menjadi tiga jenis, yaitu positif, netral, dan negatif, meskipun ada juga yang menggunakan nilai orientasi yang kontinu atau tidak diskrit.
2. *Fined - grained sentiment analysis:* Objek yang diklasifikasikan bukan pada level dokumen, melainkan pada level kalimat dalam sebuah dokumen [7].

BAB III METODOLOGI

A. Waktu dan Tempat Pelaksanaan Kerja Praktik

Pelaksanaan kerja praktik dimulai pada tanggal 3 Juni sampai 28 Juni 2024 di Kantor Perwakilan Wilayah Bank Indonesia Provinsi Lampung, Jalan Hasanuddin No. 38, Gunung Mas, Teluk Betung Utara, Bandar Lampung, Lampung 35211.

B. Diagram Alir

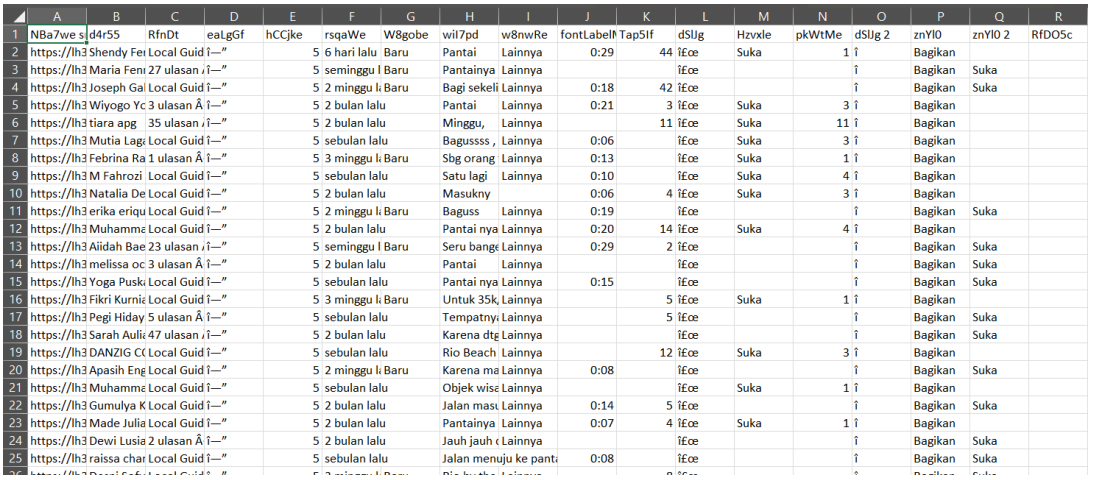


Gambar 1 Diagram Alir

C. Metode Pengolahan Data

a. Pengambilan Data

Pada penelitian ini menggunakan data primer, yaitu data *review* pengunjung ke Pantai Rio By The Beach di *Google Maps* pada tahun 2024 dengan jumlah data awal 698 dan atribut 18. Data tersebut diperoleh langsung dari *Google Maps* menggunakan metode *web scraping* dengan bantuan *extensions* bernama *Instant Data Scraper*. Web Scraping digunakan untuk melakukan ekstraksi data ulasan dan rating pengunjung ke Pantai Rio By The Beach dari *google maps* dengan tujuan melihat tingkat kepuasan pengunjung pada pantai tersebut menggunakan analisis sentimen. Hasil scraping data penelitian terlihat pada Gambar 2.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	NBa7we s	d4r55	RfnDt	eaLgGf	hCCjke	rsqaWe	W8gobe	wiI7pd	w8nwRe	fontLabelM	Tap5If	dSlJg	Hzvxle	pkWtMe	dSlJg 2	znYl0	znYl0 2	RfDO5c
2	https://lh3	Shendy Fe	Local Guid	î—”	5	6 hari lalu	Baru	Pantai	Lainnya	0:29	44	îŒ	Suka	1	î	Bagikan		
3	https://lh3	Maria Fen	27 ulasan ·	î—”	5	seminggu l	Baru	Pantainya	Lainnya			îŒ			î	Bagikan	Suka	
4	https://lh3	Joseph Ga	Local Guid	î—”	5	2 minggu l	Baru	Bagi sekeli	Lainnya	0:18	42	îŒ			î	Bagikan	Suka	
5	https://lh3	Wiyogo Yc	3 ulasan ·	î—”	5	2 bulan lalu		Pantai	Lainnya	0:21	3	îŒ	Suka	3	î	Bagikan		
6	https://lh3	tiara apg	35 ulasan ·	î—”	5	2 bulan lalu		Minggu,	Lainnya		11	îŒ	Suka	11	î	Bagikan		
7	https://lh3	Mutia Lag	Local Guid	î—”	5	sebulan lalu		Bagussss ,	Lainnya	0:06		îŒ	Suka	3	î	Bagikan		
8	https://lh3	Febrina Ra	1 ulasan ·	î—”	5	3 minggu l	Baru	Sbg orang	Lainnya	0:13		îŒ	Suka	1	î	Bagikan		
9	https://lh3	M Fahrozi	Local Guid	î—”	5	sebulan lalu		Satu lagi	Lainnya	0:10		îŒ	Suka	4	î	Bagikan		
10	https://lh3	Natalia De	Local Guid	î—”	5	2 bulan lalu		Masukny		0:06	4	îŒ	Suka	3	î	Bagikan		
11	https://lh3	erika eriqu	Local Guid	î—”	5	2 minggu l	Baru	Baguss	Lainnya	0:19		îŒ			î	Bagikan	Suka	
12	https://lh3	Muhamma	Local Guid	î—”	5	2 bulan lalu		Pantai nya	Lainnya	0:20	14	îŒ	Suka	4	î	Bagikan		
13	https://lh3	Aiidah Bae	23 ulasan ·	î—”	5	seminggu l	Baru	Seru bange	Lainnya	0:29	2	îŒ			î	Bagikan	Suka	
14	https://lh3	melissa oc	3 ulasan ·	î—”	5	2 bulan lalu		Pantai	Lainnya			îŒ			î	Bagikan	Suka	
15	https://lh3	Yoga Puska	Local Guid	î—”	5	sebulan lalu		Pantai nya	Lainnya	0:15		îŒ			î	Bagikan	Suka	
16	https://lh3	Fikri Kurnia	Local Guid	î—”	5	3 minggu l	Baru	Untuk 35k,	Lainnya		5	îŒ	Suka	1	î	Bagikan		
17	https://lh3	Pegi Hiday	5 ulasan ·	î—”	5	sebulan lalu		Tempatnya	Lainnya		5	îŒ			î	Bagikan	Suka	
18	https://lh3	Sarah Aulia	47 ulasan ·	î—”	5	2 bulan lalu		Karena dtg	Lainnya			îŒ			î	Bagikan	Suka	
19	https://lh3	DANZIG C(	Local Guid	î—”	5	sebulan lalu		Rio Beach	Lainnya		12	îŒ	Suka	3	î	Bagikan		
20	https://lh3	Apasih Eng	Local Guid	î—”	5	2 minggu l	Baru	Karena ma	Lainnya	0:08		îŒ			î	Bagikan	Suka	
21	https://lh3	Muhamma	Local Guid	î—”	5	sebulan lalu		Objek wisa	Lainnya			îŒ	Suka	1	î	Bagikan		
22	https://lh3	Gumulya K	Local Guid	î—”	5	2 bulan lalu		Jalan masu	Lainnya	0:14	5	îŒ			î	Bagikan	Suka	
23	https://lh3	Made Julia	Local Guid	î—”	5	2 bulan lalu		Pantainya	Lainnya	0:07	4	îŒ	Suka	1	î	Bagikan		
24	https://lh3	Dewi Lusia	2 ulasan ·	î—”	5	2 bulan lalu		Jauh jauh (	Lainnya			îŒ			î	Bagikan	Suka	
25	https://lh3	raissa char	Local Guid	î—”	5	sebulan lalu		Jalan menuju ke panta		0:08		îŒ			î	Bagikan	Suka	

Gambar 2 Dataset

b. *Preprocessing* Data

i. Data Cleaning

Pada tahap ini, data dibersihkan dengan menangani *missing value*, serta data yang duplikat. Jumlah data setelah dibersihkan sebanyak 350 data dengan 3 atribut.

ii. Manipulation Data

Pada penelitian ini dilakukan *manipulation data* berupa penghapusan beberapa kolom yang dianggap tidak penting pada analisis, yaitu atribut "NBa7we src", "RfnDt", "eaLgGf", "rsqaWe", "W8gobe", "w8nwRe", "fontLabelMedium", "Tap5If", "dSlJg", "Hzvxle", "pkWtMe", "dSlJg 2", "znYl0", "znYl0 2", "RfDO5c". Kemudian dilakukan perubahan nama atribut sesuai isi data, yaitu atribut "d4r55" menjadi "nama", atribut “hCCjke" menjadi "rating", dan atribut “wiI7pd" menjadi "review".

Tabel 1 Penerapan *Manipulation Data*

nama	rating	review
erika erique	5	Baguss pantainya bersih Ombaknya besar tapi masih relatif aman tapi tetap hati-hati â€¦
Maria Fenny	5	Pantainya bersih, datang kesini weekdays gak terlalu rame.. sedikit catatan : - Sign belokan setelah keluar tol kalo bisa diperbesar, sempet kebablasan krn ga lihat sign nya.. â€¦

iii. Labelling

Dalam proses analisis sentimen, langkah awal adalah memberi label pada data. Setiap ulasan dengan *rating* ≥ 3 diberi label sentimen positif, sebaliknya jika *rating* < 3 akan berlabel negatif.

Tabel 2 Penerapan *Labelling*

rating	label
5	positif
3	positif
2	negatif

iv. Case Folding

Proses selanjutnya mengubah semua huruf dalam atribut "*review*" menjadi huruf kecil (*lowercase*). Misalnya, dalam analisis sentimen ulasan pengunjung Pantai Rio By The Beach kata "Indah" dan "indah" harus dianggap sama agar tidak menghasilkan klasifikasi yang berbeda. Dengan menerapkan *case folding*, kedua kata tersebut akan diubah menjadi "indah", sehingga tidak ada perbedaan antara keduanya dalam analisis.

Tabel 3 Penerapan *Case Folding*

INPUT	OUTPUT
Pantai dan pemandangannya bagus dan bersih lagi	pantai dan pemandangannya bagus dan bersih lagi

v. Tokenizer

Tahap *tokenizer* akan mengidentifikasi kata-kata dalam teks menjadi beberapa urutan yang terpotong oleh spasi atau karakter.

Tabel 4 Penerapan *Tokenizer*

INPUT	OUTPUT
Pantai dan pemandangannya bagus dan bersih lagi	[pantai, dan, pemandangannya, bagus, dan, bersih, lagi]

vi. StopWord Removal

Stopword Removal akan menghapus kata yang tidak relevan di dalam suatu kalimat berdasarkan daftar *stopword*, yaitu "nya", "yang", "ga", "gak", "itu", "t4", "juga", "ini", "untuk".

c. Ekstraksi Fitur

i. Count Vectorizer

Tahap *Count Vectorizer* akan menghitung jumlah kemunculan masing-masing kata pada dokumen, sehingga dapat disebut juga sebagai metode *raw count*.

Tabel 5 Penerapan *Count Vectorizer*

INPUT	OUTPUT
[bagus]	(111,[2],[1.0])

ii. TF-IDF

Metode TF-IDF akan menghitung bobot setiap kata yang paling umum digunakan pada *information retrieval*.

Persamaan yang umum digunakan:

$$\text{TF} = \frac{jumlah\ kemunculan\ kata\ dalam\ dokumen}{jumlah\ kata\ total\ dalam\ dokumen} \quad (3)$$

Misalnya, jika kata "bagus" muncul 5 kali dalam sebuah komentar yang memiliki total 100 kata, maka TF untuk kata "bagus" dalam dokumen tersebut adalah $\frac{5}{100} = 0.05$.

Tabel 6 Penerapan TF-IDF

INPUT	OUTPUT
[bagus]	(111,[2],[1.29643...

d. Splitting Data

Data dibagi menjadi dua, yaitu *data train* dan *data test* dengan rasio perbandingan 60:40. Hal ini berarti 60% data akan masuk ke subset pelatihan (*train*), dan 40% data akan masuk ke subset pengujian (*test*). *Data train* digunakan untuk melatih model *machine learning*. Model akan belajar dari data ini untuk menemukan pola dan hubungan antara fitur dan label (dalam kasus ini, *rating* dan teks ulasan).

Data test digunakan untuk mengevaluasi seberapa baik model dapat melakukan prediksi terhadap data yang belum pernah dilihat sebelumnya.

e. Modelling

i. Naive Bayes

Pada pemodelan digunakan model *Naive Bayes* sebagai salah satu metode klasifikasi untuk melihat keberadaan suatu fitur tertentu dalam suatu kelas bergantung atau tidak pada keberadaan fitur lainnya. Klasifikasi ini akan menganalisis kesempatan setiap anggota kelas seperti probabilitas suatu tupel adalah milik kelas tertentu.

ii. Visualisasi *Word Cloud*

Word cloud akan menghasilkan visualisasi grafis berdasarkan frekuensi kemunculan kata dalam teks sesuai label positif dan negatif. Dalam word cloud kata-kata yang lebih sering muncul akan ditampilkan dengan ukuran lebih besar, sehingga lebih menonjol dibandingkan kata-kata lain. Semakin sering suatu kata muncul dalam dokumen, semakin besar ukuran kata tersebut dalam visualisasi.

f. Evaluasi Model

1. K-fold Cross Validation

Pada evaluasi model menggunakan metode *K-fold Cross Validation* ini digunakan k=5 sampai 10, kemudian diperoleh performa terbaik dengan akurasi maksimum sebesar 89% pada k=7. Maka, selanjutnya evaluasi model digunakan dengan k=7.

2. Confusion Matrix

Confusion Matrix akan menghitung kinerja atau tingkat kebenaran proses klasifikasi. *Confusion Matrix* berbentuk matriks persegi dengan ukuran $n \times n$, di mana n adalah jumlah kelas dalam dataset. Struktur dasar dari *Confusion Matrix* untuk model klasifikasi biner adalah sebagai berikut:

Tabel 7 *Confusion Matrix*

	Actually Positive (1)	**Actually Negative (0)**
Predicted Positive (1)	True positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negative (FNs)	True Negatives (TNs)

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

Dimana,

TP : *True Positive*, yaitu jumlah data positif yang terklasifikasi oleh sistem.

TN : *True Negative*, yaitu jumlah data bukan positif yang terklasifikasi oleh sistem.

FN : *False Negative* yaitu jumlah data negatif namun terklasifikasi salah oleh sistem.

FP : *False Negative* yaitu jumlah data positif namun terklasifikasi salah oleh sistem.

BAB IV HASIL DAN PEMBAHASAN

A. Deskripsi Data

Tabel 8 Variabel pada data *review* pengunjung Pantai Rio By The Beach

Variabel	Definisi	Deskripsi
y	Sentimen/Labelling	Sebuah penanda sentimen yang menggambarkan nilai kata apakah bersifat positif atau negatif
x_1	Review	Berisi ulasan pengunjung ke Pantai Rio By The Beach di *Google Maps*
x_2	Rating	Sebuah representasi nilai rasio tingkat kepuasan pengunjung Pantai Rio By The Beach

Data yang digunakan adalah data *review* pengunjung ke Pantai Rio By The Beach pada *Google Maps* pada tahun 2024. Variabel terikat pada penelitian ini adalah label berupa positif dan negatif (y), variabel ini akan memprediksi atau mengklasifikasikan sentimen ulasan pengunjung berdasarkan data yang tersedia. Sedangkan variabel bebas pada penelitian ini adalah *review* pengunjung (x_1), *rating* (x_2). *Review* pengunjung sebagai variabel bebas karena ulasan tersebut mengandung informasi kualitatif yang kaya tentang pengalaman dan persepsi pengunjung. Teks ulasan dianalisis untuk menentukan apakah sentimen yang terkandung di dalamnya bersifat positif atau negatif. *Rating* sebagai variabel bebas karena nilai *rating* ini merupakan indikator langsung dari kepuasan atau ketidakpuasan pengunjung. *Rating* akan memberikan petunjuk kuantitatif yang dapat dikorelasikan dengan sentimen ulasan. *Rating* yang lebih tinggi cenderung berkorelasi dengan sentimen positif, sementara *rating* yang lebih rendah cenderung berkorelasi dengan sentimen negatif.

B. Labelling

Dari 350 data yang sebelumnya sudah dilakukan cleaning, diperoleh 2 kelompok label dengan sentimen positif sebanyak 314 atau 90% dari total keseluruhan data dan 36 sentimen negatif atau sebanyak 10% dari keseluruhan data. Artinya, 90% pengunjung Pantai Rio By The Beach memberikan *review* positif pada *Google Maps*, sedangkan 10% lainnya memberikan *review* negatif.

Tabel 9 Jumlah Sentimen Pengunjung Pantai Rio By The Beach

Label	**Jumlah**
Positif	314
Negatif	36

C. Model Naive Bayes

Tabel 10 *Confusion Matrix* Model *Naive Bayes*

	1	0
1	134	12
0	15	2

Tabel *confusion matrix* di atas diperoleh nilai *accuracy* 0,83, artinya model Naive Bayes secara keseluruhan benar dalam mengklasifikasikan sentimen pengunjung 83% dari waktu. *Precision* 0,91 berarti dari semua prediksi model yang mengatakan bahwa sentimen pengunjung adalah positif, 91% di antaranya benar-benar positif. *Recall* 0,89 berarti dari semua kasus yang benar-benar positif, model berhasil mengidentifikasi 89% di antaranya.

Tabel 11 Hasil Evaluasi Model *Naive Bayes*

Classifier	***Accuracy***	***Precision***	***Recall***
Naive Bayes	0,83	0,91	0,89

D. Evaluasi Model K-fold Cross Validation

Tabel 12 *Confusion Matrix* Model *K-fold Cross Validation* dengan k=7

	1	0
1	122	6
0	9	2

Pada evaluasi model *K-fold Cross Validation* dilakukan dengan nilai k-fold dari 5 sampai 10, kemudian diperoleh nilai performa k-fold terbaik yaitu k=7. Pada tabel *confusion matrix* di atas diperoleh *accuracy* sebesar 0,89, artinya dari seluruh prediksi yang dibuat oleh model dengan k=7, 89% di antaranya adalah benar. *Precision* 0,95 berarti ketika model memprediksi bahwa sentimen pengunjung adalah positif, 95% dari waktu prediksi tersebut benar. *Recall* 0,93 berarti dari semua kasus yang sebenarnya positif, model berhasil mengidentifikasi 93% di antaranya dengan benar. Ini berarti

model K-fold Cross Validation dengan k=7 cukup baik dalam menemukan sentimen positif yang sebenarnya dalam data.

Tabel 13 Hasil Evaluasi Model *K-fold Cross Validation* dengan k=7

Classifier	***Accuracy***	***Precision***	***Recall***
K-fold Cross Validation	0,89	0,95	0,93

E. Visualisasi

a. *Word Cloud* Positif

Tampilan *Word Cloud* atau frekuensi kata-kata dari label positif yang sering muncul membahas mengenai keindahan pantai, kebersihan, pasir putih, harga tiket terjangkau, pantai bagus terlihat pada Gambar 3.

Gambar 3 *Word Cloud* Positif

b. *Word Cloud* Negatif

Tampilan *Word Cloud* atau frekuensi kata-kata dari label negatif yang sering muncul membahas mengenai akses atau jalan menuju ke pantai, parkir, fasilitas, tempat berteduh yang terbatas.

Gambar 4 *Word Cloud* Negatif

BAB V PENUTUP

A. Kesimpulan

- Berdasarkan penelitian dan pembahasan data pengunjung Pantai Rio By The Beach pada *Google Maps*, diperoleh kesimpulan bahwa pemodelan menggunakan metode *Naive Bayes* diperoleh nilai akurasi sebesar 0.8343558282208589, artinya sebanyak 83% ulasan pengunjung terklasifikasikan dengan benar ke dalam label positif dan negatif. Kemudian, pada pemodelan *K-fold Cross Validation* dengan k=7 diperoleh nilai akurasi sebesar 0. 8992805755395683, artinya model *Naive Bayes* yang digunakan dapat menjelaskan klasifikasi dari ulasan pengunjung ke dalam setiap kelas sebesar 89%.
- Dari 350 data yang sudah dilakukan *cleaning*, diperoleh 314 data atau 90% pengunjung Pantai Rio By The Beach memberikan *review* positif pada *Google Maps*, sedangkan 36 data atau 10% lainnya memberikan *review* negatif.
- Faktor utama pengunjung memberikan *review* positif yaitu mengenai keindahan pantai, kebersihan, pasir putih, harga tiket terjangkau, pantai bagus, sedangkan pada *review* negatif mengenai akses atau jalan menuju ke pantai, parkir, fasilitas, tempat berteduh yang terbatas.

B. Saran

- Dalam melakukan penelitian terhadap analisis sentimen pengunjung ke Pantai Rio By The Beach dapat dilakukan dengan metode lain untuk melihat dan membandingkan tingkat akurasi yang lebih baik dibandingkan dengan metode *Naive Bayes Classifier*, seperti menggunakan metode *Support Vector Machines* (SVM) atau *Convolutional Neural Networks* (CNN).
- Pemerintah Daerah dan Kantor Perwakilan Bank Indonesia Provinsi Lampung dapat bekerjasama dengan pihak pengelola Pantai Rio By The Beach untuk meningkatkan kualitas fasilitas, biaya dan pengelolaan parkir, akses jalan menuju pantai, serta memperbanyak tempat berteduh. Upaya ini dapat meningkatkan jumlah pengunjung ke Pantai Rio By The Beach.

DAFTAR PUSTAKA

[1] D. R. Anggraini, "DAMPAK SEKTOR PARIWISATA PADA PERTUMBUHAN EKONOMI DAERAH LAMPUNG," *Jurnal Bisnis Darmajaya,* vol. 07, no. 02, pp. 116-117, 2021.

[2] O. P. Lampung, "Dinas Kominfotik Prov. Lampung," 24 Juli 2024. [Online]. Available: https://ppid.lampungprov.go.id/detail-post/FOILA-Perkuat-Kolaborasi-Sinergi-Promosi-Investasi-dan-Perdagangan-untuk-Mencapai-Pertumbuhan-Ekonomi-Lampung-Yang-Berkelanjutan.

[3] G. A. Buntoro, "ANALISIS SENTIMEN HATESPEECH PADA TWITTER DENGAN METODE NAÏVE BAYES CLASSIFIER DAN SUPPORT VECTOR MACHINE," *Jurnal Dinamika Informatika,* vol. 5, no. 2, p. 3, 2016.

[4] T. B. A. A. E. P. Nurirwan Saputra, "ANALISIS SENTIMEN DATA PRESIDEN JOKOWI DENGAN PREPROCESSING NORMALISASI DAN STEMMING MENGGUNAKAN METODE NAIVE BAYES DAN SVM," *Jurnal Dinamika Informatika,* vol. 5, no. 1, pp. 4-5, 2015.

[5] I. SurvyanaWahyudi, " Big data analytic untuk pembuatan rekomendasi koleksi film personal menggunakan Mlib. Apache Spark," *Berkala Ilmu Perpustakaan dan Informasi,* vol. 14, no. 1, pp. 11-25, 2018.

[6] A. Imron, "ANALISIS SENTIMEN TERHADAP TEMPAT WISATA DI KABUPATEN REMBANG MENGGUNAKAN METODE NAIVE BAYES CLASSIFIER," Yogyakarta, 2019.

[7] J. C. Aponno, "Penerapan Algoritma Sentiment Analysis Dan Naïve Bayes Terhadap Opini Pengunjung Di Tempat Wisata Pantai Pintu Kota, Kota Ambon," *Jurnal Teknik Informatika dan Sistem Informasi ,* vol. 9, no. 4, pp. 180-3188, 2022.

[8] T. H. P. F. R. U. Lio Wilianto, "ANALISIS SENTIMEN TERHADAP TEMPAT WISATA DARI KOMENTAR PENGUNJUNG DENGAN MENGGUNAKAN METODE NAÏVE BAYES CLASSIFIER STUDI KASUS JAWA BARAT," *Prosiding SNATIF,* p. 439, 2017.

[9] D. T. Anggraeni, "FORECASTINGHARGA SAHAM MENGGUNAKAN METODE SIMPLE MOVING AVERAGEDAN WEB SCRAPPING," *Jurnal Ilmiah MATRIK,* vol. 21, no. 3, p. 237, 2019.

[10] F. Ratnawati, "Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter," *JURNAL INOVTEK POLBENG - SERI INFORMATIKA,* vol. 3, no. 1, p. 52, 2018.

[11] H. H. J. S. Anasthasya Averina, "Analisis Sentimen Multi-Kelas Untuk Film Berbasis Teks Ulasan Menggunakan Model Regresi Logistik," *TEKNIKA,* vol. 11, no. 2, pp. 123-128, 2022.

[12] A. A. Maarif, " PENERAPAN ALGORITMA TF-IDF UNTUK PENCARIAN KARYA ILMIAH," *Jurnal Teknik Informatika,* pp. 1-5, 2015.

[13] M. F. R. Y. M. Alfian Al Arif, "Perbandingan MetodeData Mining untuk Prediksi Curah Hujan dengan Algoritma C4.5, Naïve Bayes,dan KNN," *SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat,* p. 191, 2022.

[14] J. J. A. Limbong, "Analisis Klasifikasi Sentimen Ulasan pada E-Commerce Shopee Berbasis Word Cloud dengan Metode Naïve Bayes dan K-Nearest Neighbor," *JURNAL Fakultas Teknologi Informasi UKSW,* pp. 2-3, 2022.

[15] B. W. A. R. H. Riza Rizqi Robbi Arisandi, "APLIKASI NAÏVE BAYES CLASSIFIER(NBC)PADA KLASIFIKASI STATUS GIZI BALITA STUNTINGDENGAN PENGUJIAN K-FOLD CROSS VALIDATION," *JURNAL GAUSSIAN,* vol. 11, no. 1, pp. 130-139, 2022.

[16] A. I. G. Heni Sulastri, "PENERAPAN DATA MINING DALAM PENGELOMPOKAN PENDERITA THALASSAEMIA," *Jurnal Nasional Teknologi dan Sistem Informasi ,* vol. 03, no. 02, p. 300, 2017.

[17] Y. A. S. M. T. F. Raditya Rinandyaswara, "PEMBENTUKAN DAFTAR STOPWORDMENGGUNAKAN TERM BASED RANDOM SAMPLINGPADA ANALISIS SENTIMEN DENGAN METODE NAÏVE BAYES(STUDI KASUS: KULIAH DARING DI MASA PANDEMI)," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK),* vol. 9, no. 4, pp. 717-724, 2022.