ANALISIS SENTIMEN ULASAN TERHADAP INFRASTRUKTUR PARIWISATA DI PULAU KUMALA MENGGUNAKAN METODE NAÏVE BAYES DENGAN FITUR EKSTRAKSI BERT

Sidiq Indrajati Yusuf, Rudiman, Rofilde Hasudungan
Teknik Informatika, Universitas Muhammadiyah Kalimantan Timur
Jl. Ir. H. Juanda No.15, Sidodadi, Kec. Samarinda Ulu, Kota Samarinda, Kalimantan Timur
1911102441031@umkt.ac.id

ABSTRAK

Destinasi Wisata Pulau Kumala yang dikenal memiliki banyak daya tarik mengalami penurunan pengunjung karena kurangnya perawatan. Penelitian ini menganalisis sentimen masyarakat mengenai Pulau Wisata Kumala yang berada di Kota Tenggarong menggunakan klasifikasi Naïve Bayes dengan fitur ekstraksi BERT. Tujuan penelitian ini untuk mengetahui sentimen masyarakat tentang Wisata Pulau Kumala, dan mengevaluasi hasil model yang digunakan. Dengan metode klasifikasi Naïve Bayes dan fitur ekstraksi BERT dari 1538 data berbahasa Inggris yang didapatkan melalui google maps, penelitian ini melakukan Analisis Sentimen yang menghasilkan hasil sentimen yang dominan dari lima kelas sentimen dan hasil evaluasi model yang digunakan bersama dengan fitur seleksi chi-square. Hasil analisis sentimen menunjukkan nilai sentimen dominan pada kelas *positive*. Evaluasi hasil Naïve Bayes dengan rasio 70:30 menghasilkan akurasi sebesar 85%, dan akurasi menjadi 86% setelah menggunakan fitur seleksi chi-square.

Kata kunci : *Analisis Sentimen, Pulau Kumala, Google Maps Review, Naïve Bayes, BERT, Chi-Square*

1. PENDAHULUAN

Pulau kumala merupakan salah satu objek wisata budaya yang terletak di tengah-tengah Sungai Mahakam yang berada di Kota Tenggarong, Kabupaten Kutai Kartanegara. Pulau ini memadukan aspek modern dan budaya tradisional sebagai daya tarik wisatawan yang berkunjung dengan ikon pulau ini yaitu patung Lembuswana. Selain itu, Pulau Wisata Kumala juga memiliki beragam infrastruktur yang menarik seperti Lamin Adat, Kolam Naga, Jembatan Penyebrangan Repo-Repo, dan bangunan menarik lainnya yang hanya terdapat di Pulau Kumala [1].

Meskipun destinasi Pulau Kumala ini terkenal dan memiliki banyak daya tarik, pada tahun 2017 terdapat penurunan pengunjung hingga 27% yang salah satunya disebabkan kurangnya perawatan objek wisata, fasilitas, dan infrastruktur yang tidak optimal [2]. Pengumpulan data yang relevan dengan Pulau Kumala yang dimulai dari tahun 2017 hingga tahun 2025 diperlukan untuk mengetahui apakah penurunan pengunjung ini mempengaruhi pada pandangan sentimen Masyarakat mengenai tempat Wisata Pulau Kumala.

2. TINJAUAN PUSTAKA

Analisis sentimen yaitu proses untuk mengidentifikasi sebuah polaritas dari suatu data dalam bentuk dokumen atau teks yang hasil akhirnya akan merujuk pada emosional konotasi positif, negatif, ataupun netral [3].

Google Maps adalah sebuah layanan dalam bentuk peta digital berbasis website dan native smartphone, dengan berbagai macam fitur seperti peta dunia, informasi ulasan dari sebuah tempat, hingga fitur menentukan rute tercepat dari tempat ke tempat tertentu. Adanya Google Maps dapat membuat wisatawan Pulau Kumala untuk memberikan umpan balik mengenai hal-hal yang berkaitan terhadap pengalaman di tempat wisata dengan mudah [4].

Web Scraping yaitu teknik mengambil data dari website tertentu secara terstruktur yang akan menghasilkan data yang diperoleh sesuai dengan format program yang telah ditentukan tanpa adanya user melakukan pengambilan data secara manual [5].

Naïve Bayes merupakan salah satu metode klasifikasi untuk mengambil keputusan dengan menghitung probabilitas menggunakan teori Bayes secara keseluruhan yang kemudian akan dikombinasikan dengan nilai frekuensi [6].

Metode BERT (Bidirectional Encoder Representation from Transformers) yaitu salah satu model deep learning yang biasa digunakan untuk penerapan pemahaman bahasa. Fitur ekstraksi BERT dapat memahami relasi antar kata dengan baik melalui self-attention layer dan dapat berjalan dua arah dalam memahami token yang lalu dengan token yang akan datang [7].

3. METODE PENELITIAN

3.1. Objek Penelitian

Objek pada penelitian ini adalah data ulasan masyarakat mengenai infrastruktur Pariwisata yang ada di Pulau Kumala melalui Google Maps.

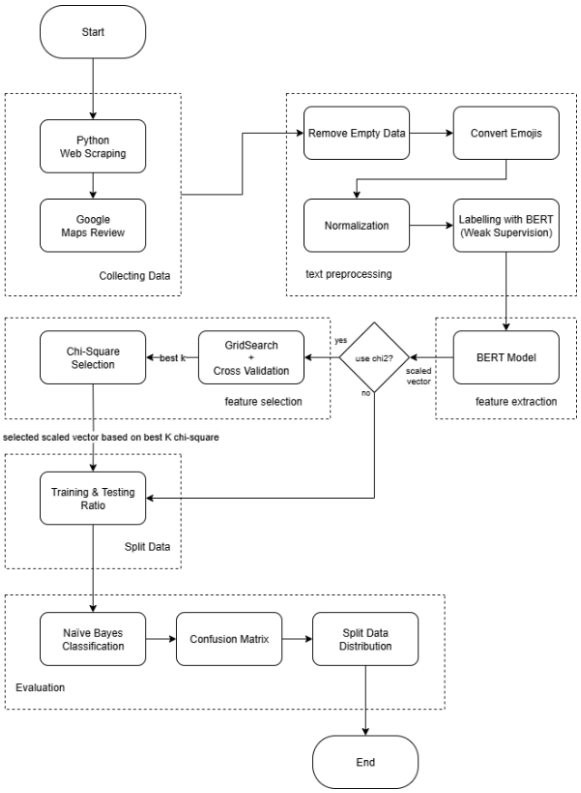
3.2. Alat dan Bahan

Penelitian ini memerlukan alat perangkat keras dan perangkat lunak pendukung guna mencapai hasil analisis sesuai tujuan penelitian. Berikut adalah detail spesifikasi perangkat yang digunakan.

a. Perangkat Keras
perangkat keras yang digunakan dengan spesifikasi perangkat prosesor AMD Ryzen 5 6600H Radeon Graphics (12CPUs, ~3.3GHz), RAM 16GB, SSD 512GB, Windows 11 64-bit
b. Perangkat Lunak
Pemrograman bahasa Python digunakan versi 3.13.x dan beberapa library dari pihak ke-3 dalam membantu menyelesaikan analisis sentimen yang akan dilakukan, seperti matplotlib yang menampilkan data dalam bentuk visual diagram, NLTK, pandas, dan library pendukung lainnya. IDE (*Integrated Development Environment*) yang digunakan yaitu *visual studio code*.

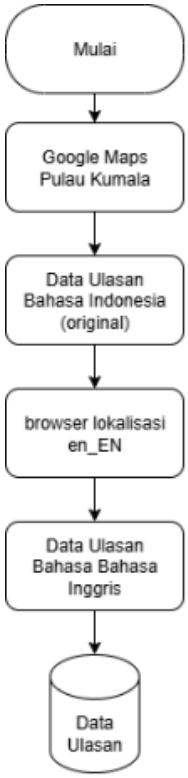
3.3. Prosedur Penelitian

Penelitian ini dilakukan dengan prosedur teknik web scraping yang bersumber dari Google Maps terhadap ulasan infrastruktur Pariwisata yang ada di Pulau Kumala.



Gambar 1. Alur penelitian

a. Pengumpulan Data
Pengumpulan data dilakukan menggunakan metode web scraping dalam melakukan ekstraksi data dari sebuah halaman web secara cepat [8]. Penelitian ini menggunakan dataset yang telah didapatkan dari google maps review terhadap infrastruktur Pariwisata yang ada di Pulau Kumala.



Gambar 2. Alur Pengumpulan Data

Penelitian ini menggunakan data ulasan Pulau Kumala yang sudah diterjemahkan otomatis oleh google kedalam Bahasa Inggris karena model pelabelan data yang akan digunakan terlatih di Bahasa Inggris. Originalitas data ulasan yaitu Bahasa Indonesia. Data yang didapatkan akan digunakan sebagai sumber data analisis sentimen.

b. Pre-Processing
Tahapan pertama sebelum melakukan analisis data yaitu melakukan persiapan data yang telah diterima dari pengumpulan data yang masih dalam bentuk mentah atau belum ada proses manipulasi apapun, agar bisa diproses dalam bentuk yang lebih terstruktur [6].

- *Remove Empty Data* merupakan proses untuk menghapus baris atau data yang tidak memiliki informasi atau kosongan.
- *Convert emojis* merupakan proses untuk melakukan konversi dari emoji menjadi teks yang memiliki arti.
- *Normalization* dilakukan untuk menghapus karakter yang tidak memiliki visual ketika ditampilkan seperti (“\t”, “\\n”, “\b”). Tahapan ini juga mengganti spasi duplikat atau baris baru menjadi spasi, dan mengganti teks emoji seperti (:smile::smile:) menjadi (smile smile).
- *Labelling Data* atau Pelabelan Data yaitu proses untuk memberikan label/anotasi secara manual terhadap data sebagai konteks agar dapat dipahami oleh bahasa mesin dan dapat menjalankan klasifikasinya [9]. Namun banyaknya data yang akan diberikan label secara manual akan memakan waktu dan energi

yang cukup banyak, maka dari itu penelitian ini menggunakan metode *weak supervision*, yaitu pemberian label melalui sampel data yang sudah di training tanpa memberikan label secara manual menggunakan *pre-trained* model yang telah di *fine-tuning* untuk analisis sentimen, yaitu model *bert-base-multilingual-uncased-sentiment* dari pustaka nlptown. Model ini adalah uncased dan mendukung sentiment review berbahasa inggris dengan total pelatihan data mencapai 150 ribu data.

c. Fitur Ekstraksi (BERT)
Fitur ekstraksi pada penelitian ini menggunakan algoritma BERT (Bidirectional Encoder Representations from Transformers) Model. BERT yaitu salah satu model deep learning pemahaman bahasa yang dapat memahami relasi antar kata dengan baik dan dapat berjalan dua arah melalui token yang lalu dan token selanjutnya [7].

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Yang dimana Nilai Q sebagai antrian (queries) dari nilai matriks, K sebagai matriks yang menampung nilai kunci (*keys*), dan V sebagai matriks yang menampung nilai dari aktual data (*values*).

d. GridSearch CV
GridSearch yaitu bagian fungsi dari modul scikit-learn yang digunakan untuk melakukan validasi dan memberikan hyperparameter secara otomatis dan sistematis, kemudian mencari parameter yang dapat memberikan hasil yang paling optimal terhadap model yang dikembangkan [10]. GridSearch CV akan digunakan untuk mencari nilai terbaik K melalui nested cross-validation yang akan digunakan sebagai nilai K pada tahapan chi-square.
Cross validation yaitu teknik yang biasa digunakan untuk memprediksi performa model melalui nilai K yang biasanya memiliki nilai antara 5 dan 10, atau dikenal sebagai out-of-sample testing [11]. Metode *cross-validation* yang digunakan: *10-fold* dirumuskan $\{\lfloor n_{features} * r \rfloor \mid r \in \{0.1,0.2,0.3,\dots,1.0\}\}$ dan refined search $k \in [K_{best} - 20, K_{best} + 21], k \in Z$.

e. Fitur Seleksi Chi-Square
Chi-Square adalah algoritma fitur seleksi yang digunakan untuk mengukur tingkat fleksibilitas antara kategori dan term. Berdasarkan penelitian sebelumnya yang pernah dilakukan menggunakan Chi-Square, menggunakan fitur seleksi dapat meningkatkan akurasi yang dihasilkan [12].

f. Split Data
Split data adalah proses membagi dataset menjadi dua bagian: (1) data training, digunakan untuk melakukan pembentukan dan penyesuaian parameter pada model, dan (2) data testing, digunakan untuk mengevaluasi hasil akurasi yang didapatkan [13]. Penelitian ini menggunakan rasio dengan hasil akurasi yang tertinggi dengan fitur seleksi dalam evaluasi hasil analisa dan fitur seleksi chi-square.

g. Evaluasi
Evaluasi yaitu tahapan yang dilakukan di akhir sebagai proses mengevaluasi mengenai kualitas model, kemampuan dan struktur kinerja dari model yang telah dikembangkan berdasarkan hasil dari prediksi yang dihasilkan [14]. Adapun evaluasi pada penelitian ini yaitu Klasifikasi Naïve Bayes, confusion matrix, dan hasil distribusi split data.
- *Naïve Bayes* sebagai klasifikasi algoritma yang menggunakan nilai maksimum dengan melakukan prinsip pengelompokan nilai estimasi ke dalam sampel nilai berdasarkan probabilitas kategori [15].
- *Confusion matrix* yaitu salah satu metode yang dapat merepresentasikan antara nilai hasil dari prediksi dengan kondisi aktual yang dihasilkan oleh algoritma [16].
- $Accuracy = \frac{Jumlah\ Prediksi\ Benar}{Total\ Data}$
- Mengevaluasi hasil distribusi split data dalam bentuk visual perbandingan jumlah data yang di *training*/*testing* dengan rasio 70:30, 80:20, 90:10 sebagai perbandingan akurasi.

4. HASIL DAN PEMBAHASAN

4.1. Pengumpulan Data

Pengumpulan data dilakukan di sebelas tempat infrastruktur yang ada di Pulau Kumala. Pengumpulan data ini dilakukan pada tanggal 6 Juli 2025 pukul 12:53AM (+0900) dengan estimasi program melakukan *scraping* di sebelas tempat selama 6 Menit 46.0 detik, termasuk program intentional delay yang dilakukan secara acak antara 1.16 detik hingga 2.31 detik pada tiap iterasi *scraping*.

Tabel 1. Daftar infrastruktur & jumlah review

No	Infrastruktur	Total	Estimasi (detik)
1	Jembatan Penyebrangan Pulau Kumala	1198	280.21
2	Wisata Tenggarong Kolam Naga	149	39.15
3	Kumala Island	68	17.13
4	Pulau Kumala	55	16.60
5	Patung Lembuswana Area SPS	31	12.47
6	Dayak Experience Center	24	10.08
7	Sky Tower Kumala	9	11.07
8	Kumala Central Park	1	4.33
9	Taman Air Mancur Pulau Kumala	1	6.16
10	Arena Balap Gokart	1	4.31
11	Candi Naga Lembuswana	1	5.09
11 Infrastruktur		**1538**	**406.06**

Dari Tabel 1 dapat dilihat terdapat 1538 data yang berhasil diperoleh dari google maps berdasarkan total dari masing-masing Infrastruktur, kemudian akan digabungkan menjadi satu dataset dari semua Infrastruktur.

Tabel 2. Hasil pengumpulan data

No	Name	Review	Stars.Value	Humanized_Timestamp
1	Fachrul Rozie F	OK, I have it	5	a year ago
2	Teras	A unique and interesting island located in the...	5	4 year ago
3	Oktoberianto Wonodiharjo	Hopefully I can walk here again	4	3 years ago

Dari Tabel 2 dapat dilihat tiga sampel hasil pengumpulan data, hanya kolom "review" saja yang akan digunakan untuk tahap *preprocessing*.

4.2. Pre-Processing

Tahapan ini akan mengolah data mentah menjadi bentuk yang konsisten dan menghindari kemungkinan error ketika melakukan pelabelan.

a. Remove Empty Reviews

Tahapan pertama yaitu menghapus data yang tidak memiliki teks review atau isi kosongan. Pada tahapan scraping sebelumnya, peneliti menggunakan indikator "[no_review]" untuk baris yang tidak memiliki review.

Tabel 3. Remove empty review

No	Sebelum	No	Sesudah
1	OK, I have it	1	OK, I have it
…			
1538	[no_review]	1355	Excellent

Berdasarkan tabel 3 terdapat 183 data yang tidak memiliki teks review telah dibersihkan dengan total jumlah data dari 1538 menjadi 1355.

b. Convert Emojis

teks ulasan yang mengandung emoji akan diubah menjadi teks yang dapat dibaca dan lebih memiliki ungkapan ekspresi.

Tabel 4. Convert emojis

Id	Sebelum	Id	Sesudah
68	Good ✿ ✿ …	68	Good:cherry_blossom::cherry_blossom: …
332	Because it's still a pandemic, the bridge isn't open yet😢 …	332	Because it's still a pandemic, the bridge isn't open yet:crying_face: …
327	The bridge is nice, the view of the Mahakam River is beautiful 🥹 …	327	The bridge is nice, the view of the Mahakam River is beautiful :smiling_face_with_hearts: …

Dari tabel 4, terdapat 87 baris data yang berubah dan 151 total emoji yang berhasil dikonversi.

c. Normalization

Teks akan dibersihkan dari karakter yang non-printable, mengandung spasi duplikat, dan mengubah teks dari (contoh. :smile: menjadi smile).

Tabel 5. Normalization

No	Sebelum	No	Sesudah
1	Good:cherry_blossom::cherry_blossom: …	1	Good(cherry_blossom cherry_blossom) …
213	This was the last destination we visited from a series of other tourist attractions in Tenggarong. I thought it would be normal. It turns o…	213	This was the last destination we visited from a series of other tourist attractions in Tenggarong.I thought it would be normal. It turns o…

Dari tabel 5 menghasilkan data yang lebih bersih dengan total perubahan mencapai 148 data.

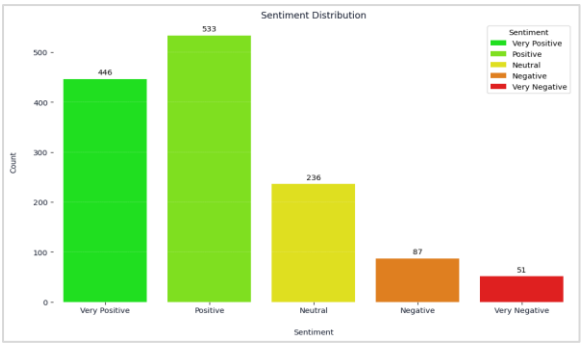
d. Labelling with BERT

Pelabelan otomatis dilakukan menggunakan model *pre-trained bert-base-multilingual-uncased-sentiment* yang telah di fine-tuned untuk sentimen analisis Bahasa Inggris. Model ini menghasilkan 5 kelas sentimen: (0) Very Negatif, (1) *negative*, (2) *neutral*, (3) *positive*, (4) *very positive*. Atribut yang digunakan sebagai pelabelan yaitu kolom "review". Berikut adalah hasil pelabelan menggunakan model ini.

Tabel 6. Labelling with bert

Id	Review	Sentiment Value
0	Okay have	2
1	The place is good for taking photos	3
27	Not properly maintained, good view, only few rest room available, 2 or 3 seater bycle rent availabe here	1

Hasil distribusi pelabelan ditampilkan pada gambar berikut.



Gambar 3. Distribusi sentimen pelabelan

Dilihat dari Gambar 3 hasil visualisasi distribusi sentimen, label *positive* memiliki nilai tertinggi, dan terendah yaitu sentimen label very negative.

e. Fitur Ekstraksi

Pada tahapan fitur ekstraksi menggunakan BERT, kolom yang digunakan yaitu review yang telah melalui proses pre-processing. Hasil dari fitur ekstraksi BERT untuk kolom review yaitu representasi token [CLS] yang memiliki dimensi 1 hingga 768 yang disimpan ke dalam kolom features.

Tabel 7. Hasil fitur ekstraksi

No	Review	Sentiment Value	Features
1	Okay have	2	[-0.20016572, 0.13576284, 0.08363976, -0.49322...
2	The place is good for taking photos	3	[-0.24281251, -0.61732364, -0.24264336, 0.0956...
3	Pleasant	3	[0.16553348, -0.21731246, 0.16979443, 0.08522...

Dari tabel 7 menghasilkan tiga sampel hasil tokenisasi yang dilakukan oleh BERT. Nilai pada kolom tersebut yaitu hasil representasi vektor dari token [CLS] awal atau hasil transformasi dari 12 layer BERT, yang dimana setiap angka adalah hasil kombinasi konteks dari BERT.

Tabel 8. Sampel token embeddings layer ke-12 pada token embeddings (5-dimensi pertama)

No	Token	Sentiment Value
1	[CLS]	[-0.24281251 -0.61732364 -0.24264336 0.09568704 0.23196378, …]
2	the	[-0.62106967 -0.28161368 -0.35939455 0.6808527 0.10839283, …]
3	place	[-0.49298033 -0.5710596 -0.14354725 0.28489292 0.6086844, …]
4	is	[-0.29613644 -0.5687712 -0.29665878 0.3388526 0.7234439, …]
5	good	[-0.7152506 -0.80630624 -0.2115109 0.07126725 0.4959701, …]
6	for	[-0.8483955 -1.0037807 -0.57002354 0.50184935 0.178357, …]
7	taking	[-0.5734824 -0.27009287 -0.0618571 -0.01218496 0.24707893, …]
8	photos	[-0.32986352 -0.19797173 -0.06783786 0.15218265 0.0853916, …]
9	[SEP]	[-0.5972064 -0.30988738 -0.25462297 0.12497092 -0.03026566, …]

Pada tabel 8, setiap token akan ditambahkan token khusus dari BERT pada awalan (yaitu [CLS]) dan akhiran (yaitu [SEP]), kemudian BERT akan memberikan *token embeddings* (*hidden state*) dengan vektor base dari BERT (dimensi 768) hasil representasi dari 12-layer transformers dari BERT.

f. GridSearch CV

Target kolom dalam mencari nilai K untuk fitur seleksi chi-square yang sebagai nilai y adalah kolom “features” yang merupakan hasil dari fitur ekstraksi BERT dengan model nlptown/*bert-base-multilingual-uncased-sentiment*. Berikut adalah hasil test menggunakan *cross-validation* untuk menemukan nilai dengan performa terbaik variabel K top fitur untuk *chi-square* sebagai initial broad search.

(todo table measurement / CV-test results)

g. Fitur Seleksi Chi-Square

Setelah didapatkan nilai terbaik K yaitu 696 melalui tahapan GridSearch CV, nilai ini akan digunakan sebagai jumlah fitur yang akan dipilih untuk fungsi *SelectKBest* parameter “k”. Hasil dari *chi-square* akan melalui tahapan proses yang sama seperti fitur ekstraksi, namun perbedaan fitur matriks yang digunakan adalah hasil nilai keluaran dari fungsi *SelectKBest*.

Tabel 9. Fitur numeriks vektor bert

No	Token	Sentiment Value
1	[CLS]	array([[0.34184784, 0.62269235, 0.47651443, ..., 0.45119017, 0.45569235, 0.3482998], [0.31299138, 0.26210666, 0.2947412 , ..., 0.62560993, 0.35538065, 0.60015744], [0.5892938 , 0.4536362 , 0.33532557, ..., 0.6901846 , 0.31188154, 0.7653103], ..., [0.20631704, 0.26779607, 0.3490469 , ..., 0.64441174, 0.30493793, 0.57998616], [0.4822041 , 0.48282605, 0.38297123, ..., 0.7187191 , 0.31490967, 0.9621338], [0.75536615, 0.38033402, 0.3655236 , ..., 0.82220614, 0.2542835 , 0.75778604]], shape=(1355, 768), dtype=float32)

Pada tabel 9, nilai *shape* 1355 adalah total jumlah data ulasan, dan nilai 768 merupakan jumlah fitur yang digunakan.

Tabel 10. Fitur numeriks vektor bert (chi-square)

No	Token	Sentiment Value
1	[CLS]	array([[0.34184784, 0.62269235, 0.3546297 , ..., 0.45119017, 0.45569235, 0.3482998], [0.31299138, 0.26210666, 0.7396266 , ..., 0.62560993, 0.35538065, 0.60015744], [0.5892938 , 0.4536362 , 0.7327851 , ..., 0.6901846 , 0.31188154, 0.7653103], ..., [0.20631704, 0.26779607, 0.6759781 , ..., 0.64441174, 0.30493793, 0.57998616], [0.4822041 , 0.48282605, 0.6273177 , ..., 0.7187191 , 0.31490967, 0.9621338], [0.75536615, 0.38033402, 0.61244494, ..., 0.82220614, 0.2542835 , 0.75778604]], shape=(1355, 696), dtype=float32)

Pada tabel 10, nilai dimensi vektor berubah menjadi 696, sesuai dari nilai best K yang didapatkan dan telah diseleksi kolom features-nya berdasarkan akurasi tertinggi dari GridSearch CV.

h. Split Data

Perbandingan split data yang digunakan yaitu 90:10, 80:20, dan 70:30. Untuk hasil evaluasi analisis sentimen, rasio yang digunakan adalah hasil dengan akurasi yang tertinggi dengan menggunakan fitur seleksi.

4.3. Evaluasi

Tahapan ini akan mengolah data mentah menjadi bentuk yang konsisten dan menghindari

a. Naïve Bayes

Hasil klasifikasi akan dibedakan berdasarkan rasio data yang didapat dari split data 90, 80, dan 70 ke dalam model Naïve Bayes yang telah dibentuk. Hasil akurasi menggunakan Naïve Bayes dengan tiap rasio dilengkapi metode penerapan fitur seleksi chi-square dan tanpa penerapan fitur seleksi dalam mendapatkan akurasi.

Tabel 11. Hasil akurasi dengan klasifikasi naive bayes

No	Ratio	Feature Dimension	Accuracy
1	70:30	Default Size = 768	0.85%
		With Chi-Square (K=696)	0.86%
2	80:20	Default Size = 768	0.83%
		With Chi-Square (K=696)	0.82%
3	90:10	Default Size = 768	0.82%
		With Chi-Square (K=696)	0.81%

Pada tabel 11, nilai akurasi tertinggi yaitu dengan rasio 70:30 dengan menggunakan fitur seleksi chi-square yang mencapai 86% dan tanpa chi-square yaitu 85%. Penelitian ini akan menggunakan rasio 70:30 sebagai tahapan evaluasi karena memiliki akurasi tertinggi dibanding rasio yang lainnya. Berikut adalah hasil tabel nilai evaluasi dari klasifikasi Naïve Bayes yang dihasilkan dengan rasio 70:30 dan fitur seleksi.

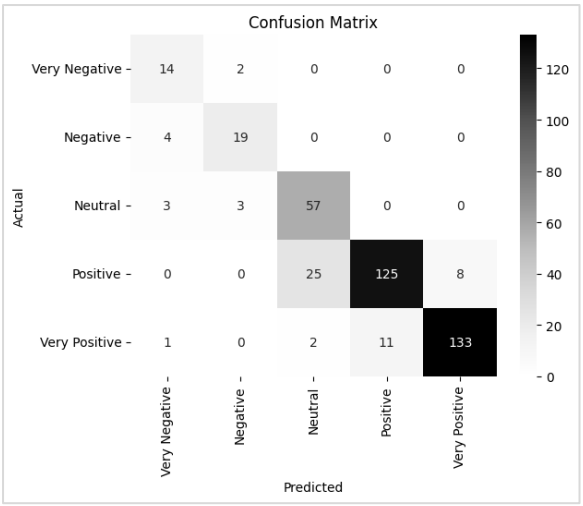
Tabel 12. Evaluasi rasio 70:30 dengan chi-square

	Precision	Recall	F1-Score
0 (very negative)	0.64	0.88	0.74
1 (negative)	0.79	0.83	0.81
2 (neutral)	0.68	0.90	0.78
3 (positive)	0.92	0.79	0.85
4 (very positive)	0.94	0.90	0.92
Accuracy			0.86
Macro Avg	0.79	0.86	0.82
Weighted Avg	0.87	0.86	0.86

Pada tabel 12, Rasio 70:30 dengan chi-square menghasilkan akurasi mencapai 86%. Nilai imbalance terjadi di kelas *very negative* dan *neutral* dibanding kelas seperti *Very Positive* yang mencapai 0.94.

b. Confusion Matrix

Hasil Akurasi dengan total data testing sebanyak 407 data (rasio 70:30) menggunakan model BERT dengan klasifikasi Naïve Bayes divisualisasikan dalam bentuk confusion matrix



Gambar 4. Confusion matrix 70:30 (chi-square)

Pada gambar 4, prediksi benar tertinggi terdapat pada label *very positive* dengan 133 dari 147 data. Berikut adalah hasil berdasarkan tiap kategori.

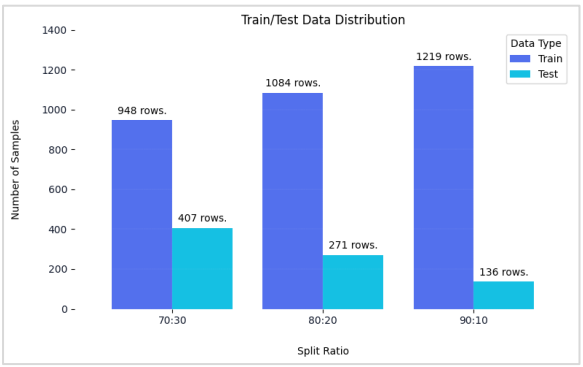
- *Very Negative*, memiliki 14 prediksi benar dan 2 tidak benar dari 16 data.
- *Negative*, memiliki 19 prediksi benar dan 4 tidak benar dari 23 data.
- *Neutral*, memiliki 57 prediksi benar dan 6 tidak benar dari 63 data.
- *Positive*, memiliki 125 prediksi benar dan 33 tidak benar dari 158 data.
- *Very Positive*, memiliki 133 prediksi benar dan 14 tidak benar dari 147 data.

Dengan perhitungan uncuk mencari akurasi:

$$Accuracy = \frac{348}{407} = 0.85503 = 0.86 = 86\%$$

c. Distribusi Training/Testing Data

Pembagian jumlah data distribusi data training dan data testing ditampilkan pada Gambar 3.3 Train/Test Data Distribution dengan total 1355 data.



Gambar 5. Training/testing data distribution

Pada gambar 5, rasio 70:30 yang digunakan untuk evaluasi terbagi menjadi 948 data training dan 407 data testing.

5. KESIMPULAN DAN SARAN

Berdasarkan pembahasan dan pengujian yang dilakukan, pengumpulan data melalui web scraping menghasilkan 1538 data dengan proses labeling menggunakan metode weak supervision menunjukkan nilai yang cenderung kuat untuk sentimen positif. Klasifikasi Naïve Bayes dapat memberikan hasil yang cukup dan akurat dari dataset sentimen berbahasa Inggris yang telah melalui teks preproses, meskipun terdapat *label imbalance* (kelas kecil seperti *very negative* dan *neutral*), hasil akurasi model yang cukup tinggi mencapai 85%. Dengan menggunakan chi-square, akurasi meningkat 1% menjadi 86% dengan 696 fitur yang digunakan dengan rasio 70:30. Untuk meningkatkan performa dari BERT, disarankan untuk mencoba menggunakan metode yang mendukung nilai *dense vector representation* seperti Support Vector Machine (SVM) atau Logistic Regression. Pendekatan fitur seleksi yang lain seperti PCA (Principal Component Analysis), t-SNE (t-distributed Stochastic Neighbor Embedding) yang juga mendukung nilai *dense vector* yang dihasilkan oleh BERT. Hasil pembobotan label kelas yang imbalance dapat dilakukan pendekatan seperti resampling atau class weighting.

DAFTAR PUSTAKA

[1] B. L. Pratiwi, "Pengelolaan Daya Tarik Wisata Pulau Kumala oleh Dinas Pariwisata Kabupaten Kutai Kartanegara," *Jurnal Administrasi Bisnis Fisipol Unmul*, vol. 8, no. 1, pp. 46–54, 2020.

[2] E. D. Arsita and N. S. S. Giriwati, "Tingkat Kepuasan Pengunjung Terhadap Objek Wisata Pulau Kumala di Kutai Kartanegara," *RUAS*, vol. 20, no. 2, pp. 97–108, 2022.

[3] Merinda Lestandy, Abdurrahim Abdurrahim, and Lailis Syafa'ah, "Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naïve Bayes," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 4, pp. 802–808, Aug. 2021, doi: 10.29207/resti.v5i4.3308.

[4] E. Elinda, H. Yuliansyah, and M. I. A. Latiffi, "Sentiment Analysis of the Sheikh Zayed Grand Mosque's Visitor Reviews on Google Maps Using the VADER Method," *International Journal of Advances in Data and Information Systems*, vol. 5, no. 1, pp. 71–84, Apr. 2024, doi: 10.59395/ijadis.v5i1.1320.

[5] D. Rudini, D. G. Purnama, and A. A. Khan, "Penggunaan Teknik Web Scraping dalam Aplikasi Pengambilan Data dari Google Maps untuk Menunjang Digital Marketing," *Lentera: Multidisciplinary Studies*, vol. 2, no. 1, pp. 10–19, Nov. 2023, doi: 10.57096/lentera.v2i1.61.

[6] A. Munawaroh, R. Ridhoi, and R. Rudiman, "SENTIMENT ANALYSIS DENGAN NAÏVE BAYES BERBASIS ORANGE TERHADAP RESIKO PEMBANGUNAN IKN," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 1, pp.

587–592, Feb. 2024, doi: 10.36040/jati.v8i1.8454.

[7] A. S. Yazid and E. Winarko, "Fine-Tuning BERT untuk Menangani Ambiguitas Pada POS Tagging Bahasa Indonesia," *Jurnal Linguistik Komputasional (JLK)*, vol. 6, no. 2, pp. 57–64, Sep. 2023, doi: 10.26418/jlk.v6i2.148.

[8] V. Pichiyan, S. Muthulingam, S. G, S. Nalajala, A. Ch, and M. N. Das, "Web Scraping using Natural Language Processing: Exploiting Unstructured Text for Data Extraction and Analysis," *Procedia Comput Sci*, vol. 230, pp. 193–202, 2023, doi: https://doi.org/10.1016/j.procs.2023.12.074.

[9] Tamanna, "Text Labeling Tools for NLP," Medium. Accessed: Jul. 09, 2025. [Online]. Available: https://medium.com/@tam.tamanna18/text-labeling-tools-for-nlp-a-comprehensive-overview-79b7ba26e98a

[10] R. Muzayanah, D. A. A. Pertiwi, M. Ali, and M. A. Muslim, "Comparison of gridsearchcv and bayesian hyperparameter optimization in random forest algorithm for diabetes prediction," *Journal of Soft Computing Exploration*, vol. 5, no. 1, pp. 86–91, Apr. 2024, doi: 10.52465/joscex.v5i1.308.

[11] I. K. Nti, O. Nyarko-Boateng, J. Aning, and others, "Performance of machine learning algorithms with different K values in K-fold cross-validation," *International Journal of Information Technology and Computer Science*, vol. 13, no. 6, pp. 61–71, 2021.

[12] Y. D. Kirana and S. Al Faraby, "Sentiment analysis of beauty product reviews using the K-nearest neighbor (KNN) and TF-IDF methods with chi-square feature selection," *Journal of Data Science and Its Applications*, vol. 4, no. 1, pp. 31–42, 2021.

[13] V. R. Joseph and A. Vakayil, "SPlit: An Optimal Method for Data Splitting," *Technometrics*, vol. 64, no. 2, pp. 166–176, Apr. 2022, doi: 10.1080/00401706.2021.1921037.

[14] R. Merdiansah, S. Siska, and A. Ali Ridha, "Analisis Sentimen Pengguna X Indonesia Terkait Kendaraan Listrik Menggunakan IndoBERT," *Jurnal Ilmu Komputer dan Sistem Informasi (JIKOMSI)*, vol. 7, no. 1, pp. 221–228, Mar. 2024, doi: 10.55338/jikomsi.v7i1.2895.

[15] H. Chen, S. Hu, R. Hua, and X. Zhao, "Improved naive Bayes classification algorithm for traffic risk management," *EURASIP J Adv Signal Process*, vol. 2021, no. 1, p. 30, Dec. 2021, doi: 10.1186/s13634-021-00742-6.

[16] Resika Arthana, "Mengenal Accuracy, Precision, Recall dan Specificity serta yang diprioritaskan dalam Machine Learning," https://rey1024.medium.com/mengenal-accuracy-precission-recall-dan-specificity-serta-yang-diprioritaskan-b79ff4d77de8