# KLASIFIKASI TEKS ULASAN PADA WEB TRIPADVISOR TENTANG WISATA ALAM PULAU LOMBOK MENGGUNAKAN METODE NAIVE BAYES CLASSIFIER

*Tripadvisor Web Review Text Classification Of Lombok Island Nature Tourism Using Naive Bayes Classifier Method*

Yuni Oktaviani, I Gde Putu Wirarama Wedashwara W.* , Ariyan Zubaidi
Dept Informatics Engineering, Mataram University
Jl. Majapahit 62, Mataram, Lombok NTB, INDONESIA
*Email: yunioktaviani121@gmail.com, wirarama@unram.ac.id, zubaidi13@unram.ac.id*

***Abstract***

*The tourism sector has also been greatly affected by advances in technology and the internet, many positive impacts are felt in the tourism sector. TripAdvisor is the world's largest travel platform that helps travelers optimize the potential of every trip, reviews on TripAdvisor contain various aspects of assessment and are in the form of mixed sentiment (such as positive and negative). For this reason, it is necessary to have a TripAdvisor review classification on tourist attractions so that it can be seen what aspects of the assessment are often discussed by visitors and can determine a specific assessment whether the review is positive or negative. This study uses the Naïve Bayes Classifier method. This model is very fast in training. and can be used on smaller datasets. Although it is a simple model, this method is able to produce a fairly high accuracy. In this study, the authors also use the additional selection feature of mutual information, MI is used to measure the contribution of a term to the right category based on presence or absence. the classification carried out using the nave Bayes classifier method with additional feature selection obtained an accuracy value of 85.86%, while without using mutual information feature selection and obtaining an accuracy of 83.80%, based on the two classifications that have been carried out, the accuracy difference is 2.06%, which means that the classification process using mutual information feature selection can increase the accuracy by 2.06%. The presented research has contributed as text mining system for specific topic about Lombok Island tourism that refered by tripadvisor.*

***Keywords:*** *TripAdvisor, Mutual Information, Wisata Lombok, Naïve Bayes Classifier*

**Penulis Korespondensi*

## 1. PENDAHULUAN

Indonesia diakui sebagai salah satu negeri yang elok dengan berbagai keindahan alamnya baik di darat atau di laut. Di Indonesia terdapat banyak pantai, gua, laguna, estuari, hutan mangrove, padang lamun, rumput laut, dan terumbu karang yang menghiasi negeri ini[1]. Pulau Lombok merupakan salah satu pulau dengan destinasi wisata yang cukup terkenal di Indonesia karena keindahan alam dan banyaknya objek wisata seperti wisata alam, wisata budaya dan wisata kuliner.

Bidang pariwisata saat ini juga sudah sangat terpengaruh oleh kemajuan teknologi dan internet, banyak dampak positif yang dirasakan dibidang pariwisata, dengan adanya kemajuan teknologi dan internet, sekarang banyak situs online yang menawarkan jasa booking hotel, restoran, penginapan, booking tiket transportasi seperti kereta api, pesawat, dan lainnya. Bahkan untuk orang yang sekedar ingin mencari referensi untuk rencana wisatanya, kini sudah dapat memanfaatkan banyak media sosial, untuk mencari foto-foto hingga ulasan atau review tentang pariwisata tersebut dari para pengunjung sebelumnya. Media sosial yang sering digunakan untuk acuan pariwisata adalah TripAdvisor.

TripAdvisor adalah platform wisata terbesar di dunia yang membantu wisatawan mengoptimalkan potensi setiap perjalanan, menjangkau rata-rata 390 juta pengunjung unik setiap bulannya, serta menampilkan 500 juta ulasan dan opini tentang 6,8 juta akomodasi, restoran, dan objek wisata[3]. Pengguna TripAdvisor memberikan ulasan, komentar dan penilaian (rating) pada tempat wisata. ulasan pada TripAdvisor mengandung berbagai aspek

penilaian dan berupa mixed sentiment (seperti positif dan negatif). Untuk itu, perlu adanya Klasifikasi ulasan TripAdvisor pada tempat wisata sehingga dapat diketahui aspek penilaian apa saja yang sering dibahas oleh para pengunjung dan dapat menentukan penilaian secara spesifik apakah ulasan tersebut positif atau negatif. Salah satu model yang digunakan dalam melakukan klasifikasi yaitu Naïve Bayes Classifier. Model ini sangat cepat dalam pelatihan dan dapat digunakan pada dataset yang lebih kecil. Meskipun merupakan model yang sederhana, namun metode ini mampu menghasilkan akurasi yang cukup tinggi. Kemudahan implementasi juga merupakan keuntungan besar dari Naive Bayes Classifier

Berdasarkan paparan di atas penulis mengajukan sebuah penelitian untuk merancang sebuah model untuk mengklasifikasikan ulasan pada web Tripadvisor tentang wisata alam Pulau Lombok. Penelitian ini menggunakan metode Naive Bayes Classifier. Selain metode naïve bayes classifer, diterapkan juga seleksi fitur yaitu dengan seleksi fitur Mutual information, seleksi fitur mutual.

## 2. TINJAUAN PUSTAKA

### 2.1 Tinjauan Pustaka

Penelitian mengenai pengklasifikasian suatu kalimat menggunakan metode Naive Bayes Classifier sudah pernah dilakukan oleh beberapa peneliti dalam kurun waktu 5 tahun belakangan ini. Penelitian-penelitian sebelumnya akan dijadikan sebagai rujukan ketika pelaksanaan penelitian ini.

Analisis Sentimen Tripadvisor Terhadap Pariwisata Gunung Bromo dan Gunung Semeru[6]. Penelitian yang menggunakan metode Support Vector Machine dengan data review berbahasa indonesia. Data yang digunakan berjumlah 320 review yang di dapatkan dari proses crawling menggunakan script R. Data klasifikasi dibagi menjadi 3 kondisi yaitu positif, negatif dan netral tanpa melakukan seleksi fitur. Hasil klasifikasi yang di peroleh memiliki tingkat akurasi 71% untuk gunung bromo dan 62% untuk gunung semeru.

Analisis Sentimen Terhadap Dampak Covid-19 Pada Performa Tokopedia Menggunakan Support Vector Machine [7]. Penelitian ini menggunakan metode Support vector Machine dengan menggunakan ulasan berbahasa indonesia dengan jumlah 746 ulasan sebelum covid dan 1.243 setelah covid. Data dikumpulkan dengan melakukan Web Scrapping pada situs TripAdvisor tanpa menggunakan seleksi fitur dan menghasilkan akurasi sebesar 87% untuk data sebelum covid dan 84% untuk data setelah covid.

Analisis Sentimen Opini Publik Bahasa Indonesia Terhadap Wisata Tmii Menggunakan Naïve Bayes Dan Pso [8]. Penelitian ini menggunakan metode Naive bayes Classifier dan PSO. Pada penelitan ini menggunakan data berbahsa indonesia sebanyak 50 ulasan negatif dan 50 ulasan positif yang dikumpulkan secara manual. Dimana hasil klasifikasi dengan metode naive bayes sebesar 70% namun jika diambah dengan PSO (Particle Swarm Optimization) akurasinya menjadi sebesar 94,02%.

Analisis Sentimen Pelanggan Toko Online Jd. Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi[9]. Penelitian ini menggunakan metode Naive bayes classifier dengan menambahakan konversi ikon emosi (emoticon) serta pembobotan dengan TF-IDF tanpa menggunakan seleksi fitur. Nilai akurasi yang di dapatkan sebesar 96,44%, dan mampu meningkat hingga 98% jika disertai konversi emoticon dan pembobotan dengan TF-IDF.

Analisis Sentimen Menggunakan Metode Naïve Bayes Classifier Dengan Seleksi Fitur Chi Square Dalam proses pelabelan, pemilihan fitur digunakan dan dilakukan dengan pemilihan fitur chi-square, untuk mengurangi gangguan (noise) dalam klasifikasi. Hasil penelitian menunjukkan bahwa frekuensi kemunculan fitur yang diharapkan dalam kategori benar dan dalam kategori salah memiliki peran penting dalam pemilihan fitur chi-square. Kemudian klasifikasi dengan classifier Naïve Bayes memperoleh akurasi 83% dan rata-rata harmonik 90,713%.

Telah dilakukan pula penelitian tentang sentiment analisys di jejaring sosial Twitter menggunakan algoritma naïve Bayes dengan seleksi fitur mutual information[4]. Data yang digunakan adalah sejumlah 500 tweet berbahasa inggris. Data dikelompokan ke dalam 2 kategori yaitu sentimen positif dan sentimen negatif. Akurasi yang didapatkan melalui pengujian 10-fold cross validation adalah 96.2% tanpa menggunakan seleksi fitur mutual information dan 97.9% dengan menggunakan seleksi fitur mutual information.

Berdasarkan penelitian yang telah dilakukan sebelumnya, dapat disimpulkn bahwa penggunaan naive bayes classifier tanpa menambahkan future selection menghasilkan akurasi yang kurang tinggi sedangkan dengan penambahan future selection salah satnya (mutual information) dapat menabah akurasi. Oleh karena itu, penelitian untuk klasifikasi ulasan wisata alam Pulau Lombok dapat dilakukan dengan menggunakan kedua metode tersebut serta pengambilan data dengan teknik web scraping yang

dapat mempermudah pengambilan data dalam jumlah yang cukup banyak.

Penelitian berkontribusi sebagai system terpadu dalam menambang informasi tentang topik pariwisata di pulau Lombok dengan sumber data tripadvisor. Sistem yang dibangun memiliki kontribusi sebagai data crawler, text pre-proccesing, hingga klasifikasi untuk analisis sentiment terhadap objek wisata tersebut.

### 2.2 Dasar Teori

Teori-teori dasar atau umum yang digunakan dalam penelitian ini yaitu sebagai berikut

#### 2.2.1 Ulasan

Teks ulasan adalah suatu teks yang berisi ulasan, penilaian atau review terhadap suatu karya seperti film, drama, atau sebuah buku. Ulasan juga dapat disebut review. Ulasan pada umumnya ditulis dalam bentuk artikel, sehingga teks ulasan juga dapat disebut artikel ulasan, Ulasan merupakan teks yang berfungsi untuk menimbang, menilai, dan mengajukan kritik terhadap karya atau peristiwa yang diulas tersebut[11].

#### 2.2.2 Klasifikasi teks

Klasifikasi teks adalah mengklasifikasi teks berdasarkan kategori yang sudah di tentukan. Klasifikasi teks dapat digunakan untuk mengatur, menyusun, dan mengkategorikan hampir semua hal. Misalnya, artikel baru dapat diatur berdasarkan topik, percakapan obrolan dapat diatur berdasarkan bahasa, penyebutan merek dapat diatur berdasarkan sentimen, dan sebagainya[12].

#### 2.2.3 Web *Scrapping*

web scraping adalah praktik mengumpulkan data melalui cara apa pun selain program yang berinteraksi dengan API (atau, tentu saja, melalui manusia menggunakan browser web). Ini paling sering dilakukan dengan menulis program otomatis yang menanyakan server web, meminta data (biasanya dalam bentuk HTML dan file lain yang terdiri dari halaman web), dan kemudian mem-parsing data tersebut untuk mengekstrak informasi yang diperlukan

#### 2.2.4 *Preprocessing* Teks

Preprocessing teks adalah proses pembersihan dan penyiapan teks untuk klasifikasi. Teks online biasanya berisi banyak noise dan bagian yang tidak informatif seperti tag HTML, skrip, dan iklan. Selain itu, pada level kata, banyak kata dalam teks tidak berdampak pada orientasi umum teks. Menjaga kata-kata tersebut membuat dimensi masalah menjadi tinggi dan karenanya klasifikasi lebih sulit karena setiap kata dalam teks diperlakukan sebagai satu dimensi. Ada beberapa tahap dalam preprocessing teks, seperti :

a. *Tokenization*
   *Tokenization* mengacu pada sigmentasi teks yang dipisahkan oleh spasi dan tanda baca. diproses dengan cara memisahkan setiap kata, dilakukan penghapusan special karakter lalu diubah menjadi bentuk *lower case* hingga dihasilkan list token.
b. *Stemming*
   Mengacu pada pemetaan setiap token yang dihasilkan dari proses *tokenization* menjadi bentuk jamak. *Stemming* biasanya berlaku pada kata benda, kata sifat dan kata kerja.
c. *Stop-word removal*
   *Stop-word removal* adalah *penghapusan stop-word* yang hanya berfungsi secara tata bahasa dari list token ataupun hasil dari tahap *Stemming* yang di sebut kata stemmed. Kata stemmed adalah kata-kata gramatikal yang mana tidak relevan dengan konten teks, sehingga perlu dihapus agar lebih efisien [13].

#### 2.2.5 *Feature Selection*

Seleksi fitur dapat membuat proses klasifikasi menjadi lebih efisien dan efektif dengan mengurangi jumlah data yang dianalisis, maupun mengidentifikasi fitur yang sesuai untuk dipertimbangkan dalam proses pembelajaran. Ada dua jenis metode seleksi fitur yaitu Wrappers dan Filter, Wrappers menggunakan akurasi klasifikasi beberapa algoritma sebagai fungsi evaluasinya. Wrappers harus menguji klasifikasi setiap fitur bagian yang dievaluasi filters melakukan seleksi fitur menggunakan fitur yang dipilih. Salah satu contoh metode filetrs adalah Mutual Information.

#### 2.2.6 Naive bayes classifier

Naive bayes classifier merupakan model pembelajaran probabilitas sederhana dan dapat diimplementasikan dengan sangat efisien dengan kompleksitas linier. NB adalah salah satu pengklasifikasi yang paling banyak digunakan dan memiliki beberapa properti yang membuatnya sangat berguna dan akurat. Pada saat klasifikasi, model ini akan menghasilkan kategori atau kelas yang paling tinggi probabilitasnya (VMAP) [14]. Rumus VMAP dapat dinotasikan pada persamaan 2.1

$$V_{MAP} = \underset{V\ j\ e\ V}{\arg\max} \prod_{i=1}^{n} p(X_i | V_j) P(V_j)$$

di mana :

$Vj$ \= Kategori ulasan yaitu ulasan positif dan ulasan negatif.

($Xi \mid Vj$) = Probabilitas $Xi$ pada kategori $Vj$
($Vj$) = Probabilitas dari $V$

### 2.2.7 Mutual Information

Mutual Information merupakan salah satu metode seleksi fitur yang menunjukkan seberapa banyak informasi ada atau tidaknya sebuah term memberikan kontribusi dalam membuat keputusan klasifikasi secara benar atau salah [14].

$$I(U;C) = \frac{N_{11}}{N}\log_2\frac{NN_{11}}{N_1.N_{.1}} + \frac{N_{01}}{N}\log_2\frac{NN_{01}}{N_0.N_{.1}} + \frac{N_{10}}{N}\log_2\frac{NN_{10}}{N_1.N_{.0}} + \frac{N_{00}}{N}\log_2\frac{NN_{00}}{N_0.N_{.0}}$$

di mana :

N = Jumlah ulasan yang memiliki *et* dan *e*atau N = N00 + N01 + N10 + N11.
N1. = Jumlah ulasan yang memiliki *et* atau N1. = N10 + N11.
N.1 = Jumlah ulasan yang memiliki *e*atau N.1 = N01 + N11.
N0. = Jumlah ulasan yang tidak memiliki *et* atau N0.= N01 + N00.
N.0 = Jumlah ulasan yang tidak memiliki *e*atau N.0 = N10 + N00.

Tabel 2. 1 Tabel kontingensi seleksi fitur MI

| | $ec$ = 1 | $ec$ = 0 |
|---|---|---|
| $et$ = 1 | $N11$ | $N10$ |
| $et$ = 0 | $N01$ | $N00$ |

### 2.2.8 Validasi dan Evaluasi Hasil

Validasi keakuratan sebuah model yang dibangun dapat dilakukan dengan cross validation. Salah satu metode cross validation yang populer adalah K-fold cross validation..

Selain itu, dalam mengevaluasi kinerja klasifikasi, perlu digunakan akurasi untuk mengetahui seberapa bagus model tersebut dalam melakukan klasifikasi yang diinginkan. Nilai akurasi merepresentasikan seberapa banyak keseluruhan dokumen diklasifiksikan dengan benar[15]. Rumus untuk mendapatkan nilai akurasi dapat dinotasikan pada persamaan 2.4.

$$\text{Akurasi} = \frac{\text{Total dokumen yang di klasifikasi dengan benar}}{\text{Total Dokumen}}$$

*Confusion Matrix* merupakan matrik evaluasi yang paling sering digunakan pada kasus klasifikasi teks [15]. Dalam melakukan pengukuran biasanya dibangun confusion matrix yang merupakan sebuah tabel yang terdiri atas banyaknya baris data uji yang diprediksi benar dan tidak benar oleh model klasifikasi sebagaimana yang ditunjukkan pada Tabel 2.2.

Tabel 2. 2 Tabel confusion matrix

| Data Class | Classified as positive | Classified as negative |
|---|---|---|
| Positif | true positive (tp) | false negative (fn) |
| Negative | false positive (fp) | true negative (tn) |

Dimana :
TP = Kelas yang diprediksi positif dan benar.
TN = Kelas yang diprediksi negatif dan benar.
FP = Kelas yang diprediksi positif dan salah.
FN = Kelas yang diprediksi negatif dan salah.
Rumus untuk mendapatkan nilai precision dan recall dapat dinotasikan dengan persamaan.

$$\text{Precision Positive} = \frac{tp}{tp + fp}$$

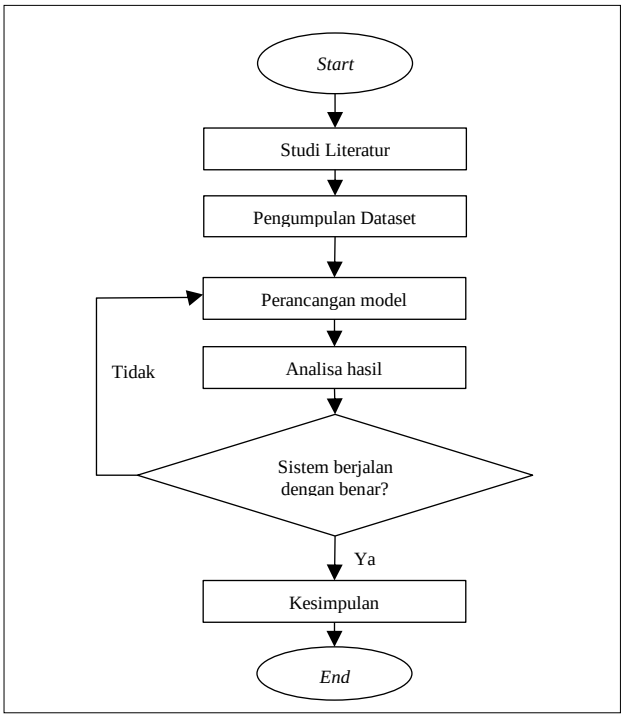
$$\text{Precision Negative} = \frac{tn}{tn + fn}$$

$$\text{Recall Positive} = \frac{tp}{tp + fn}$$

$$\text{Recall Negative} = \frac{tn}{tn + fp}$$

## 3. METODE PENELITIAN

Dalam Penelitian ini, terdapat tahapan-tahapan kegiatan yang dilakukan pada alur penelitian sebagai berikut.



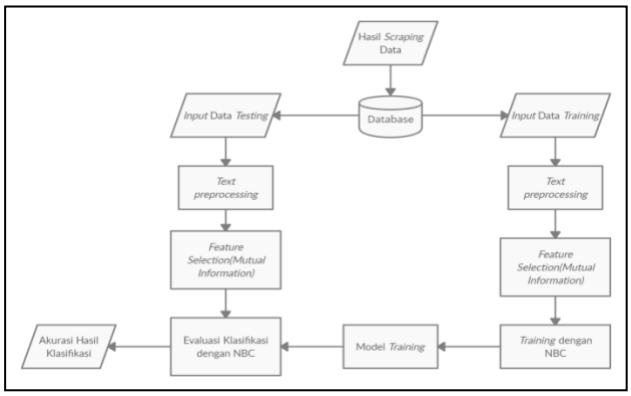
Gambar 1. Diagram alur penelitisn

Tahapan- penelitian ini terdiri dari enam tahap dimulai dari studi literature sampai dengan penarikan kesimpulan..

### 3.1 Studi LIteratur

untuk mendukung penelitian adalah mempelajari lteratur seperti buku elektronik, jurnal – jurnal penelitian, serta berbagai sumber lainnya yang berkaitan dengan materi tentang web scraping, teks processing, klasifikasi teks, metode naive bayes classifier dan penelitian sejenis yang menggunakan metode naive bayes classifier.

### 3.2 Rancangan Alur Sistem




Gambar 3. 1 Alur sistem

### 3.3 Pengumpulan data

Pengumpulan data dilakukan dengan mengumpulkan ulasan dari situs www.Tripadvisor.co.id dengan cara web Scraping menggunakan bahasa pemrograman phyton. Sraping dilakukan pada setiap objek wisata alam yang telah di tentukan. Scraping dilakukan berdasarkan URL dari masing masing objek wisata, Tag yang dimasukkan untuk mengambil ulasan adalah tag untuk judul ulasan dan tag untuk mengambil isi dari ulasan. ulasan yang di ambil hanya ulasan yang berbahasa indonesia, kumpulan ulasan yang dihasilkan akan disimpan kedalam bentuk file dengan format csv.

### 3.4 Input data training dan testing

Data training dan data testing yang digunakan dalam penelitian ini diambil dari www.TripAdvisor.co.id dengan cara Web Scraping, data yang diambil berupa ulasan dari pengguna situs untuk wisata alam yang ada di pulau lombok. Data ulasan di ambil dari wisata alam pantai, air terjun, gunung, dan pulau setelah itu data yang di ambil dibagi menjadi data testing dan data training.

### 3.1 *Text Preprocessing*

#### 3.1.1 *Tokenization*

Setelah data di dapatkan, setelahnya dilakukan proses tokenisasi untuk memisahkan setiap suku kata, menghapus spesial karakter dan mengubah setiap kata kedalam bentuk *lower case.*

#### 3.1.2 *Stop-word removal*

Setelah melalui tahap tokenisation, data akan melaui proses *Stop-word removal* yaitu pemilihan kata-kata, dimana kata-kata yang tidak berarti atau tidak memiliki makna akan dihapuskan. Proses *Stop-word removal* dilakukan menggunakan library NTLK. Berikut merupakan contoh data yang telah melaui proses *stop-word removal* setelah tokenisasi..

### 3.2 *Feature Selection with Mutual Information*

Pada *feature selection* Nilai informasi setiap fitur akan dihitung kemudian dipilih nilai tertinggi dari beberapa fitur yang akan menjadi hasil seleksi fitur. MI digunakan dalam mengukur kontribusi sebuah *term* terhadap kategori yang tepat berdasarkan kehadiran atau ketidakhadiran.

### 3.3 Klasifikasi dengan NBC

Proses klasifikasi dengan NBC terdiri dari dua tahap. Tahap pertama yaitu proses *training* data atau pelatihan dan klasifikasi. Dalam melakuikan *training* data latih ada beberapa hal yang harus di lakukan, yang pertama yaitu menghitung probabilitas kelas $p(Ci)$ kemudian menghitung probabilitas fitur.
Persamaan menghitung probabilitas kelas dapat dilihat pada persamaan :

$$p(Ci) = \frac{fd(Ci)}{|D|}$$

$fd(Ci)$ = Jumlah ulasan yang termasuk kelas $Ci$
$|D|$ = Jumlah data latih

Persamaan menghitung probabilitas kelas dapat dilihat pada persamaan :

$$p(\mathrm{Wk}|\mathrm{Ci}) = \frac{f(W|Ci)+1}{f(Ci)+|W|}$$

$f(Wk|Ci)$ = Nilai kemunculan fitur pada kelas $Ci$
$f(Ci)$ = Jumlah fitur pada kelas $Ci$
$|W|$ = Jumlah keseluruhan dari fitur (tanpa *duplicate* fitur)

### 3.4 Validasi dan Evaluasi

Validasi dilakukan dengan mennggunakan *K-fold cross validation* yang merupakan salah satu teknik validasi silang dengan cara membagi data menjadi *k* bagian dengan ukuran yang sama. Pelatihan dan pengujian dilakukan sebanyak *k* kali. Pada percobaan pertama, *subset* S1 diberlakukan sebagai data pengujian, dan *subset* lainnya digunakan sebagai data *training*. Pada percobaan ke-2, *subset* S2 diberlakukan sebagai data pengujian, kemudian *subset* lainnya digunakan sebagai data *training*. Proses

ini dilakukan sampai *k* kali dimana *subset* Sk dijadikan data pengujian[14]. pengukuran akurasi dilakukan dengan *confusion matrix.Confusion matrix* merupakan sebuah tabel yang berisi hasil data *testing* yang diprediksi benar dan salah oleh model klasifikasi. Tujuan dari *confusion matrix* yaitu melihat kinerja dari model klasifikasi[4].

## 4. HASIL DAN PEMBAHASAN

### 4.1 Data

| | Unnamed: 0 | review_text | label |
|---|---|---|---|
| 0 | 0 | tempatnya sejuk dingin natural suasananya dike... | 1 |
| 1 | 1 | bepergian benang kelambu membutuhkan stamina f... | 1 |
| 2 | 2 | musim penghujan moment terbaik berkunjung ini ... | 1 |
| 3 | 3 | harga korang asing bisa air terjun indah harga... | 1 |
| 4 | 4 | lokasinya pintu masuk gunung rinjani debit air... | 1 |
| ... | ... | ... | ... |
| 1289 | 1289 | pantai indah bersih ombak besarbagus beraktivi... | 1 |
| 1290 | 1290 | pantai indahpenginapan beragamrestaurant berag... | 0 |
| 1291 | 1291 | gili trawangan pantai biru bening pasir putihn... | 0 |
| 1292 | 1292 | gili terawangan indah cari penginapan murah ha... | 1 |
| 1293 | 1293 | pulau indah lombok utara dikelilingin pasir pu... | 1 |

[1294 rows x 3 columns]

Gambar 4.1 Kumpulan Data

### 4.2 Pelabelan Data

Ada tiga cara pelabelan dalam analisis sentimen ulasan, yaitu berdasarkan nilai rating, berdasarkan kamus lexicon dan pemberian label secara manual dengan membaca ulasan satu-persatu. Pada penelitian ini penulis menggunakan pelabelan berdasarkan kamus *lexicon*.

| | review_text | label |
|---|---|---|
| 0 | tempatnya sejuk dingin natural suasananya dike... | 1 |
| 1 | bepergian benang kelambu membutuhkan stamina f... | 1 |
| 2 | musim penghujan moment terbaik berkunjung ini ... | 1 |
| 3 | harga korang asing bisa air terjun indah harga... | 1 |
| 4 | lokasinya pintu masuk gunung rinjani debit air... | 1 |

Gambar 4.2 Data yang belum dilabeli

| | review_text | label |
|---|---|---|
| 0 | tempatnya sejuk dingin natural suasananya dike... | 1 |
| 1 | bepergian benang kelambu membutuhkan stamina f... | 1 |
| 2 | musim penghujan moment terbaik berkunjung ini ... | 1 |
| 3 | harga korang asing bisa air terjun indah harga... | 1 |
| 4 | lokasinya pintu masuk gunung rinjani debit air... | 1 |

Gambar 4.3 Data yang sudah dilabeli

### 4.3 Text Preprocessing

Data teks perlu dibersihkan sebelum masuk ke model machine learning, proses pembersihan ini disebut dengan *text preprocessing*. Terdapat beberapa tahapan dalam proses text preprocessing. Tahap *text preprocessing* secara lengkapnya akan dibahas disub-sub bab tersendiri seperti berikut.

1. *Case Folding*

Tahap ini akan mengubah semua kata menjadi huruf kecil karena tidak semua teks konsisten dalam menggunakan huruf besar sehingga perlu dirubah kedalam bentuk standar yaitu *lowercase* atau huruf kecil. Hasil *case folding* dapat dilihat pada Gambar 4.2

| | Unnamed: 0 | review_text | label |
|---|---|---|---|
| 0 | 0 | tempatnya sejuk dingin natural suasananya dike... | 1 |
| 1 | 1 | bepergian benang kelambu membutuhkan stamina f... | 1 |
| 2 | 2 | musim penghujan moment terbaik berkunjung ini ... | 1 |
| 3 | 3 | harga korang asing bisa air terjun indah harga... | 1 |
| 4 | 4 | lokasinya pintu masuk gunung rinjani debit air... | 1 |

Gambar 4.4 Hasil *case folding*

2. *Cleaning*

Tahap ini akan membuang semua kata berupa URL, *hastag*, *username* maupun email dari ulasan karena dianggap sebagai kata yang tidak tidak efektif dan tidak memiliki arti di mana untuk menganalisis sentimen tidak diperlukan banyaknya nama pengguna yang berkomentar ataupun rujukan pada suatu situs melainkan hanyalah komentar dari pengguna. Hasil *cleaning* dapat dilihat pada Gambar 4.1

| | Unnamed: 0 | review_text | label |
|---|---|---|---|
| 0 | 0 | tempatnya sejuk dingin natural suasananya dike... | 1 |
| 1 | 1 | bepergian benang kelambu membutuhkan stamina f... | 1 |
| 2 | 2 | musim penghujan moment terbaik berkunjung ini ... | 1 |
| 3 | 3 | harga korang asing bisa air terjun indah harga... | 1 |
| 4 | 4 | lokasinya pintu masuk gunung rinjani debit air... | 1 |

Gambar 4.5 Hasil *cleaning*

*3. Tokenization*

Tahap ini akan memecah teks menjadi kata atau token berdasarkan karakter pemisah yang memisahkannya seperti whitespace. Selain itu semua karakter yang bukan huruf seperti angka dan tanda baca serta delimiter lainnnya akan dihapus. Hasil *tokenization* dapat dilihat pada Gambar 4.4

```
Tokenizing Result :

0    [tempatnya, sejuk, dingin, natural, suasananya...
1    [bepergian, benang, kelambu, membutuhkan, stam...
2    [musim, penghujan, moment, terbaik, berkunjung...
3    [harga, korang, asing, bisa, air, terjun, inda...
4    [lokasinya, pintu, masuk, gunung, rinjani, deb...
Name: tokens, dtype: object
```

Gambar 4.6 Hasil *Tokenization*

4. Normalisasi

Pada proses normalisasi dilakukan proses penggantian kata dari kata tidak baku menjadi kata baku, pada proses ini kata-kata asing juga diubah dan diterjemahkan kedalam bahasa Indonesia, pengubahan data dilakukan berdasarkan dataset yang sudah dibuat sebelumnya. Hasil data yang sudah di normalisasi dapat dilihat pada Gambar 4.7

```
0    [tempatnya, sejuk, dingin, natural, suasananya...
1    [bepergian, benang, kelambu, membutuhkan, stam...
2    [musim, penghujan, moment, terbaik, berkunjung...
3    [harga, korang, asing, bisa, air, terjun, inda...
4    [lokasinya, pintu, masuk, gunung, rinjani, deb...
5    [air, terjun, lumayan, bagus, lombok, ukuran, ...
6    [salah, destinasi, wisata, alam, kawasan, kabu...
7    [air, terjun, persiapkan, stamina, tenaga, ext...
8    [air, terjun, keren, menempuh, perjalanan, sul...
9    [lombok, memiliki, pantai, indah tapi, lombok,...
Name: normalisasi, dtype: object
```

Gambar 4.7 Hasil normalisasi

5. Stopword

Tahap ini akan menghapus kata-kata umum yang tidak penting dan muncul secara berulang serta tidak memiliki makna yang berarti seperti kata sambung maupun kata ganti. proses ini dapat dilakukan dengan menyimpan daftar kata yang menurut peneliti adalah stopwords. Pada saat ini akan memperlakukan stopwords sebagai kata yang tidak mengandung makna apa pun, dan akan dihapus dalam data.. Hasil stopword dapat dilihat pada Gambar 4.8.

```
0       [tempatnya, sejuk, dingin, natural, suasananya...
1       [bepergian, benang, kelambu, membutuhkan, stam...
2       [musim, penghujan, moment, terbaik, berkunjung...
3       [harga, korang, asing, air, terjun, indah, har...
4       [lokasinya, pintu, masuk, gunung, rinjani, deb...
                              ...
1289    [pantai, indah, bersih, ombak, besar bagus, be...
1290    [pantai, indahpenginapan, beragam restaurant, ...
1291    [gili, trawangan, pantai, biru, bening, pasir,...
1292    [gili, terawangan, indah, cari, penginapan, mu...
1293    [pulau, indah, lombok, utara, dikelilingin, pa...
Name: tokens_WSW, Length: 1294, dtype: object
```

Gambar 4.8 Hasil *stopword removal*

6. *Stemming*

Tahap ini akan mengubah kata yang memiliki imbuhan menjadi kata dasar. *Stemming* berfungsi untuk menghilangkan variasi-variasi morfologi yang melekat pada sebuah kata. Dalam melakukan *stemming* digunakan NLP. Hasil *stemming* dapat dilihat pada Gambar 4.9

```
5322
------------------------
tempatnya : tempat
sejuk : sejuk
dingin : dingin
natural : natural
suasananya : suasana
dikelilingi : keliling
bukit : bukit
susana : susana
seru : seru
pelacakan : lacak
bepergian : pergi
benang : benang
kelambu : kelambu
membutuhkan : butuh
stamina : stamina
fit : fit
menikmati : nikmat
```

Gambar 4.9 Hail *stemming*

## 4.4 TF-IDF

TF-IDF adalah pembobotan dimana bobot sebuah kata berdasarkan frekuensi dokumen terbalik. Yang berarti jika sebuah kata semakin banyak muncul pada banyak dokumen, maka kata tersebut memiliki bobot yang lebih kecil[16]. Hasil pembobotan kata secara *TFIDF*, dapat dilihat pada gambar 4.12 dan 4.13 berikut.

| term | TF | TF-IDF |
|---|---|---|
| lokasi | 0.07142857142857142 | 0.16197526250785244 |
| kota | 0.07142857142857142 | 0.19618949051887485 |
| mataram | 0.07142857142857142 | 0.1988190600276403 |
| bandara | 0.07142857142857142 | 0.2114857754049914 |
| tagih | 0.07142857142857142 | 0.3343276303766318 |
| jalan | 0.07142857142857142 | 0.08403800415507016 |
| air | 0.07142857142857142 | 0.079697144273492 |
| terjun | 0.07142857142857142 | 0.1525039681191722 |
| tempuh | 0.07142857142857142 | 0.200627474883661 |
| istilah | 0.07142857142857142 | 0.4623104496072072 |
| sakit | 0.14285714285714285 | 0.6572205882284727 |
| senang | 0.14285714285714285 | 0.4068579087892193 |

Gambar 4.10 Hasil Pembobotan TF-IDF

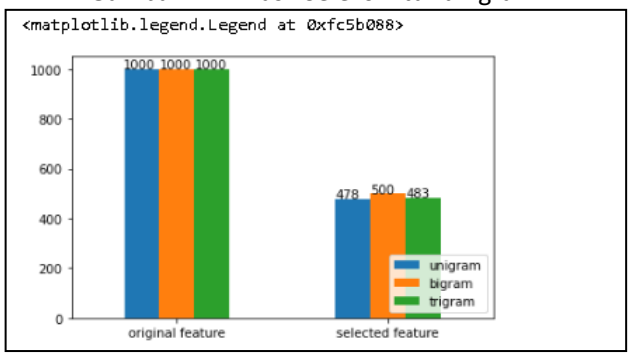
Berdasarkan Gambar 4.10, data yang paling banyak muncul dalam dataset secara berurutan yaitu air, jalan, terjun, lokasi, kota, mataram, tempuh, bandara, tagih, senang, istilah, dan sakit.

## 4.5 Seleksi Fitur

Pada seleksi fitur dilakukan pemilihan kemunculan kata terbanyak yang selanjutnya akan diurutkan sesuai rangking kemunculan kata terbanyak seperti pada Gambar 4.11, selksi fitur berguna untuk meminimalisir penggunaaan kata dalam sistem sehingga proses klasifikasi berjalan lebih efisien dan dapat mempersingkat waktu klasifikasi

| | term | rank | MI_unigram |
|---|---|---|---|
| **614** | pantai | 123.348538 | 0.358541 |
| **300** | indah | 58.693288 | 0.046401 |
| **8** | air | 56.686947 | 0.000000 |
| **680** | pulau | 55.307748 | 0.260469 |
| **489** | lombok | 54.012928 | 0.830463 |
| **248** | gili | 53.419207 | 0.000000 |
| **620** | pasir | 47.929443 | 0.000000 |
| **318** | jalan | 43.469087 | 0.000000 |
| **445** | kuta | 39.951871 | 0.000000 |
| **610** | pandang | 38.977803 | 0.047309 |

Gambar 4.11 Hasil Seleksi fitur unigram




Gambar 4.12 Grafik perbandingan setelah selekksi fitur

Grafik diatas merupakan gambaran dari perbandingan jumlah fitur sebelum dan sesudah fitur diseleksi, dengan permisalan dari 1000 fitur yang ada setelah di seleksi hanya 478 fitur yang akan digunakan dalam klasifikasi

### 4.6 Data Training dan Testing

```
# Selanjutnya kita split dataset menjadi data training dan data testing
from sklearn.model_selection import train_test_split

# Kita simpan sentiment (class) ke dalam sebuah variabel
classes=tags
X_train, X_test, y_train, y_test = train_test_split(X, classes, test_size = 0.7, random_state=30)

Feature Extraction selesai
```

Gambar 4.13 Grafik perbandingan setelah selekksi fitur

### 4.7 Klasifikasi

Sesuai dengan hasil kajian pustaka pada bab ii, penelitian ini menggunakan algoritma klasifikasi yaitu Naïve Bayes. Dalam memilih model dengan performansi paling baik dan sesuai dengan dataset maka dilakukan serangkaian percobaan yaitu menemukan jumlah data optimal,. Hasil yang diperoleh dari serangkaian percobaan tersebut yaitu sebagai berikut.

```
from sklearn import model_selection, naive_bayes
from sklearn.metrics import accuracy_score
Naivebayes = naive_bayes.MultinomialNB()
Naivebayes.fit(X_train,y_train)
# predict the labels on validation dataset
predictions_NB = Naivebayes.predict(X_test) # Use accuracy_score function to get the accuracy
print("Naive Bayes Accuracy Score ->",accuracy_score(predictions_NB, y_test)*100)

Naive Bayes Accuracy Score -> 85.8611825192802
```

Gambar 4.14 Akurasi klasifikasi dengan NBC menggunakan seleksi fitur MI

```
The model is trained on the full development set.
The scores are computed on the full evaluation set.

              precision    recall  f1-score   support

           0     0.2143    0.0545    0.0870        55
           1     0.8613    0.9671    0.9111       334

    accuracy                         0.8380       389
   macro avg     0.5378    0.5108    0.4990       389
weighted avg     0.7698    0.8380    0.7946       389
```

Gambar 4.15 Akurasi klasifikasi dengan NBC Tanpa menggunakan seleksi fitur MI

Berdasarkan hasil klasifikasi yang dilakukan dengan metode naïve bayes classifier menggunakan nilai akurasi yang didapatkan adalah sebesar 85,86 %. Dilakukan juga klasifikasi dengan metode naïve bayes classifier tanpa menggunakan seleksi fitur mutual information dan memperoleh akurasi sebesar 83,80%, berdasarkan kedua klasifikasi yang telah dilakukan, diperoleh selisih akurasi sebesar 2.06%, yang artinya proses klasifikasi dengan menggunakan seleksi fitur mutual information dapat menaikkan akurasi sebesar 2,06%

### 4.8 Evaluasi dan validasi

Evaluasi berfungsi untuk mengetahui akurasi dari model algoritma yang diusulkan. Validasi digunakan untuk melihat perbandingan hasil akurasi dari model yang digunakan dengan hasil yang telah ada sebelumnya. Teknik validasi yang digunakan adalah 10-Fold Cross Validation. Pada saat melakukan validasi, urutan dari kumpulan dokumen yang ada akan diacak. Hal ini bertujuan untuk menghindari adanya pengelompokkan dokumen yang berasal dari ketegori tertentu. Hasil pengujian masing-masing dapat dilihat pada tabel dibawah ini.

```
from sklearn.naive_bayes import MultinomialNB
from sklearn.model_selection import cross_val_score
clasfc= MultinomialNB()
scores = cross_val_score(clasfc,X_train, y_train, cv=10)

scores

array([0.84615385, 0.84615385, 0.84615385, 0.83516484, 0.83516484,
       0.84444444, 0.84444444, 0.84444444, 0.84444444, 0.84444444])
```

Tabel 4.16 Hasil 10-Cross Validation

Nilai akurasi yang dihasilkan pada setiap fold memiliki hasil yang bervariatif. Akurasi tertinggi pada fold-1, 2 dan 3 dengan akurasi yaitu 0,84. Akurasi terendah pada fold-5 yaitu dengan akurasi 0,83. Perbedaan akurasi pada setiap fold disebabkan karena karakteristik data uji dan data latih pada setiap fold berbeda-beda. Pada fold-1,2 dan 3 menghasilkan akurasi tinggi karena pada data uji dan data latih memiliki kedekatan karakteristik, sedangkan pada fold-1 antara data latih dan data uji memiliki karakteristik yang cukup jauh. Rata-rata akurasi yang diperoleh yakni 0.850463.

```
scores.mean()

0.8431013431013431
```

Tabel 4.17 Rata-rata akurasi

Tabel 4.1 Hasil Confusion Matrix

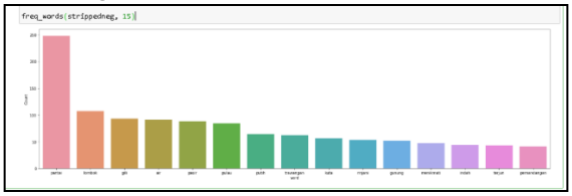
| Data Aktual | Prediksi | |
|---|---|---|
| | Negatif | Positif |
| Negatif | 0 | 55 |
| Positif | 1 | 333 |
| Akurasi | 86% | |

Berdasar Tabel 4.9 pada data ulasan berbahasa Inggris dapat diketahui bahwa, terdapat 333 data positif yang benar terprediksi masuk kedalam kelas sentimen positif dan 1 data positif yang terprediksi masuk kedalam kelas sentimen negatif, serta tidak ada data negatif yang benar terprediksi masuk kedalam kelas sentimen negatif dan ada 55 data negatif yang terprediksi masuk kedalam kelas sentiment positif.

### 4.9 Visualisasi

Dibawah ini merupakan gambar visualisasi untuk kelas positif dan kelas negative:

1. Kelas negative

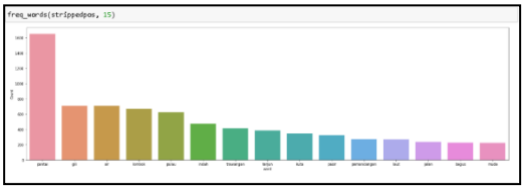


Gambar 4.18 Frekuensi Kata pada Kelas Negatif

Berdasar Gambar 4.18 diperoleh informasi bahwa pada kelas sentimen negatif kata yang paling sering muncul adalah kata-kata pantai, Lombok gili, air, pasir, pulau, putih, trawangan, kuta, rinjani, gunung, menikmati, indah, terjun, pemandangan . Kata-kata tersebut ditampilkan dalam visualisasi word cloud seperti berikut, dengan hasil yang ditampilkan pada Gambar 4.19


Gambar 4.19 Word Cloud Kelas Negatif

2. Kelas Positive



Gambar 4.20 Frekuensi Kata pada Kelas Positif

Berdasar Gambar 4.20 diperoleh informasi bahwa pada kelas sentimen positif kata yang paling sering muncul adalah kata-kata pantai,gili, air, Lombok, pulau, indah, trawangan, terjun, kuta, pasir, laut, jalan, bagus dan muda. Kata-kata tersebut ditampilkan dalam visualisasi word cloud seperti berikut, dengan hasil yang ditampilkan pada Gambar 4.21


Gambar 4.19 Word Cloud Kelas Positif

## 5. KESIMPULAN DAN SARAN

Berdasarkan Penelitian yang telah dilakukan tentang klasifikasi ulasan pada web tripadvisor tentang wisata alam pulau Lombok menggunakan metode naïve bayes, diperoleh kesimpulan bahwa :

1. Akurasi klasifikasi dengan menggunakan metode Naive bayes Classifier dalam mengklasifikasikan ulasan wisata alam pada pulau Lombok ke dalam kelas positif dan kelas negative yaitu sebesar 85,86% menggunakan seleksi fitur MI dan 83,80% tanpa menggunakan seleksi fitur MI. Oleh karena itu dengan adanya penggunaan seleksi fitur MI dapat meningkatkan akurasi sebesar 2.06%.
2. Dari klasifikasi ulasan pada data ulasan pengunjung wisata alam pada pulau Lombok dapat di di ketahui bahwa ulasan/*review* pengunjung lebih banyak yang bersentimen positif.

Adapun beberapa saran yang dapat digunakan untuk pengembangan kedepannya yaitu sebagai berikut :

1. Apabila dilakukan penelitian berikutnya disarankan menggunaan flatform yang lainnya.
2. Apabila dilakukan penelitian sebaiknya dilakukan juga pelabelan dengan cara manual agar dapat dilihat perbandingan pada hasil klasifikasi.
3. Dalam penelitian selanjutny Apabila dilakukan penelitian a disarankan menggunakan lebih banyak objek wisata yang ada di pulau Lombok

## 6. DAFTAR PUSTAKA

[1] A. J. G. Djou, “Pengembangan 24 destinasi wisata bahari kabupaten ende,” *Univ. Flores*, vol. 3, no. 1, 2013.

[2] D. S. Kusumo. D. Koesumaningrum, A. Herdiani, “Analisis Sentimen Ulasan TripAdvisor Pada Tempat Wisata Menggunakan Ontology Supported Polarity Mining (OSPM) (Studi Kasus: Bandung),” *Univ. Telkom*, pp. 1–2, 2018.

[3] I. Y. Insani. Saragih and G. I. Bhaskara, “Pencitraan Sosial Media : Studi Kasus Ulasan Tripadvisor Terhadap 5 Restaurant Terbaik Di Bali,” *Univesiitas Udayana*, vol. 7, no. 2, pp. 231–238, 2019.

[4] A. Y. Husodo. M. A. Ulfa, B. Irmawati, “Twitter Sentiment Analysis using Na¨ıve Bayes Classifier with Mutual Information Feature Selection,” *J. Comput. Sci. Informatics Eng.*, vol. 2, no. 2, pp. 106–111, 2018.

[5] D. F. Setiawan and A. Hijrian i, “Aplikasi Web Scraping Deskripsi Produk,” vol. 14, no. 1, pp. 41–47, 2020.

[6] E. Widodo *et al,* “Analisis Sentimen Tripadvisor Terhadap Pariwisata Gunung Bromo dan Gunung Semeru,” *Semin. …*, no. November, pp. 43–48, 2019, [Online]. Available: http://papersmai.mercubuana-

yogya.ac.id/index.php/smai/article/download/33/29.

[7] D. T. Wisudawati *et al.*, "Analisis sentimen terhadap dampak covid-19 pada performa tokopedia menggunakan support vector machine," pp. 87–96, 2020.

[8] R. Y. Hayuningtyas and R. Sari, "Analisis Sentimen Opini Publik Bahasa Indonesia Terhadap Wisata Tmii Menggunakan Naïve Bayes Dan Pso," *J. Techno Nusa Mandiri*, vol. 16, no. 1, pp. 37–42, 2019, doi: 10.33480/techno.v16i1.115.

[9] F.N. Sari, A. Wibowo, "Analisis Sentimen Pelanggan Toko Online Jd. Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 2, no. 2, pp. 681–686, 2019.

[10] T. Jo, *Text Mining*, vol. 36, no. 2. 2019.

[11] B. Syahid, "Pengertian Teks Ulasan, Contoh, Ciri, Tujuan, Struktur Dan Kaidahnya," no. Tersedia di https://www.gurupendidikan.co.id/teks-ulasan/, p. diakses 02-09-2020.

[12] C. Rahayu, "Klasifikasi Teks," no. Tersedia di https://mti.binus.ac.id/2020/09/03/klasifikasi-teks/, p. diakses 08-09-2020.

[13] J. Žižka, F. Dařena, and A. Svoboda, *Text Mining with Machine Learning*. 2019.

[14] A. Sabrani, I. G. W. W. Wirawan., and F. Bimantoro, "Multinomial Naïve Bayes untuk Klasifikasi Artikel Online tentang Gempa di Indonesia," *J. Teknol. Informasi, Komputer, dan Apl. (JTIKA )*, vol. 2, no. 1, pp. 89–100, 2020, doi: 10.29303/jtika.v2i1.87.