Perbandingan Performa Algoritma *XGBoost* dan *Logistic Regression* dalam Klasifikasi Sentimen Ulasan Hotel

Athallah Anargya Yogapranata
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
athallahanargya@student.telkomuniversity.ac.id

Azka Nadhira
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
azkanadhiraa@student.telkomuniversity.ac.id

Cetrin Azahra
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
cetrinazahra@student.telkomuniversity.ac.id

Abstrak

Penelitian ini bertujuan untuk menganalisis sentimen dari ulasan hotel dengan menggunakan algoritma XGBoost dan Logistic Regression. Dataset yang digunakan berisi 20.492 ulasan hotel dari TripAdvisor. Proses preprocessing mencakup pembersihan teks, tokenisasi, lematisasi, dan vektorisasi menggunakan TF-IDF. Ketidakseimbangan data ditangani dengan metode SMOTE. Evaluasi kinerja model dilakukan menggunakan confusion matrix, yang menunjukkan bahwa XGBoost memiliki akurasi 91%, precision 0.92, recall 0.91, dan F1-score 0.92, sementara Logistic Regression memiliki akurasi 90%, precision 0.93, recall 0.90, dan F1-score 0.91. Hasil penelitian menunjukkan bahwa kedua algoritma bekerja sangat baik dalam mengklasifikasikan ulasan hotel menjadi sentimen positif atau negatif.

Kata kunci: klasifikasi sentimen, logistic regression, ulasan hotel, xgboost

Abstract

This study aims to analyze hotel reviews' sentiment using XGBoost and Logistic Regression algorithms. The dataset used consists of 20,492 hotel reviews from TripAdvisor. The preprocessing steps include text cleaning, tokenization, lemmatization, and vectorization using TF-IDF. Data imbalance is addressed using the SMOTE method. Model performance evaluation is conducted using a confusion matrix, which shows that XGBoost achieves an accuracy of 91%, precision of 0.92, recall of 0.91, and F1-score of 0.92, while Logistic Regression achieves an accuracy of 90%, precision of 0.93, recall of 0.90, and F1-score of 0.91. The results indicate that both algorithms perform very well in classifying hotel reviews into positive or negative sentiments.

***Keywords: sentiment classification,* logistic regression, hotel reviews, xgboost**

I. PENDAHULUAN

Ulasan hotel dapat membantu dalam memahami sentimen pelanggan terhadap layanan, fasilitas, dan pengalaman mereka di hotel. Ulasan positif dapat digunakan dalam strategi pemasaran untuk meningkatkan citra hotel dan menarik lebih banyak pelanggan. Sebaliknya, ulasan negatif dapat digunakan sebagai peluang untuk memperbaiki citra suatu hotel dan memperbaiki kelemahan yang ditemukan. Dalam hal ini, analisis sentimen dapat memberikan pemahaman mengenai kinerja hotel secara keseluruhan berdasarkan ulasan pelanggan, sehingga dapat membantu manajemen hotel dalam mengevaluasi keberhasilan strategi pelayanan dan perbaikan fasilitas. Dalam analisis sentimen, ada beberapa algoritma pembelajaran mesin yang dikemukakan oleh banyak peneliti, diantaranya *Multinomial Naive Bayes, Random Forest, Logictic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN),* dan *Extra Trees Classifier*.

Penelitian serupa terkait analisis sentimen pernah dilakukan diantaranya yaitu seperti pada penelitian oleh A.H. Hasugian[1] dalam penelitiannya menggunakan metode *Naïve Bayes* dalam klasifikasi sentimen terhadap *review* pengguna *E-Commerce*, menghasilkan *accuracy* sebesar 99,5%, *precision* sebesar 99,49%, *recall* sebesar 100%. Pada penelitian ini kami akan menganalisis sentimen dari ulasan hotel, yang berfokus dalam perbandingan performa algoritma yang digunakan yaitu, *XGBoost* dan *Logistic Regression* dengan melakukan beberapa eksperimen pengujian dari model yang digunakan terhadap data ulasan hotel.[2] Selain itu, kami ingin mengetahui faktor-faktor yang memengaruhi ulasan hotel dan kemudian akan kami prediksi sentimen dari ulasan hotel apakah termasuk kedalam ulasan positif atau negatif.

II. STUDI LITERATUR

A. *Sentiment Analysis*

Analisis sentimen adalah teknik pemrosesan bahasa alami (NLP) untuk menganalisis emosi dan opini dari suatu teks dengan tujuan untuk memahami sikap,

penilaian, dan perasaan yang diungkapkan dalam teks tersebut[3]. Dengan analisis sentimen, dapat menentukan apakah teks tersebut termasuk sentimen positif atau negatif terhadap aspek tertentu dalam suatu kalimat. Manfaat dari analisis sentimen adalah dapat membantu mengidentifikasi area untuk perbaikan, membuat pemasaran yang lebih bertarget, dan memahami opini pelanggan mengenai produk atau layanan tertentu.

B. Tokenisasi

Tokenisasi adalah proses memecah teks menjadi unit-unit yang lebih kecil yang disebut token[4]. Tokenisasi dapat digunakan untuk analisis sentimen dan klasifikasi teks.

C. Lematisasi

Lematisasi adalah proses mengubah kata menjadi bentuk dasarnya dan tidak memiliki imbuhan[5]. Tujuan dari lematisasi untuk menghilangkan variasi bentuk kata seperti pada di kamus.

D. TF-IDF

TF-IDF merupakan singkatan dari *Term Frequency-Inverse Document Frequency* adalah sebuah statistik yang umum digunakan dalam bidang *information retrieval* (IR) dan *natural language processing* (NLP). TF-IDF berfungsi untuk mengukur pentingnya sebuah kata (*term*) dengan memberikan bobot terhadap suatu dokumen dalam sebuah kumpulan dokumen (*corpus*). Metode ini menggabungkan dua konsep untuk menghitung bobot yaitu dengan *term frequency* adalah frekuensi kemunculan kata pada kalimat dan *document frequency* adalah banyaknya kalimat yang memunculkan kata[6].

E. Algoritma *XGBoost*

XGBoost singkatan dari *Extreme Gradient Boosting*, adalah algoritma *machine learning* yang populer dan efektif untuk analisis dan klasifikasi regresi yang didasarkan pada *Gradient Boosting Decision Tree (*GBDT)[7]. Pada metode ini, dibangun model baru sehingga dapat mengurangi kesalahan dari model sebelumnya. *XGBoost* dapat menyelesaikan berbagai tugas seperti klasifikasi suata data positif atau negatif dan regresi dari suatu data.

F. Algoritma *Logistic Regression*

Logistic Regression adalah analisis regresi di mana variabel respon hanya memiliki dua kemungkinan nilai[8]. *Logistic Regression* menggunakan fungsi matematika yang disebut fungsi sigmoid untuk memetakan hubungan antara variabel independen (variabel prediktor) dan variabel dependen (variabel target). Variabel dependen pada *Logistic Regression* berbentuk kategori[8].

$$P(x) = \frac{1}{(1+ e^{-x})} \quad (1)$$

G. *Confusion Matrix*

Tabel yang digunakan untuk mengevaluasi performa model klasifikasi dalam pembelajaran mesin untuk menyatakan klasifikasi data uji yang benar dan salah[9]. Matriks ini dapat memahami seberapa baik model tersebut dapat mengklasifikasikan data dengan benar, dengan menampilkan hasil prediksi model dibandingkan dengan label sebenarnya dari data. Berikut rumus *confusion matrix* untuk menghitung *accuracy, precision,* dan *recall*[9]:

$$accuracy = \frac{TP+TN}{Total} \quad (2)$$

$$precision = \frac{TP}{TP+FP} \quad (3)$$

$$recall = \frac{TP}{TP+FN} \quad (4)$$

III. METODOLOGI PENELITIAN

A. *Dataset*

Dataset yang digunakan dalam penelitian ini adalah "*Hotel Reviews Sentiment*" yang didapatkan dari Kaggle. *Dataset* ini berisi 20.492 ulasan hotel dari Web Tripadvisor. Setiap ulasan disertai dengan teks ulasan dan *rating* dari hotel tersebut, dengan skala (1 - 5). Berikut tautan *dataset* kami, Hotel Reviews Sentiment prediction(kaggle.com).

B. *Preprocessing*

Preprocessing meliputi beberapa tahap antara lain pembersihan teks, pelabelan data, menghitung persentase tanda baca, tokenisasi, dan lematisasi.

1. Pembersihan Teks

Menghilangkan karakter non-alfabet dan mengubah semua teks menjadi huruf kecil dengan menggunakan *library reguler expression*, hal ini dikarenakan karakter non-alfabet seperti angka, tanda baca, dan simbol lainnya tidak banyak berkontribusi dalam analisis teks. Selain itu dengan menghapus karakter asing dapat memastikan bahwa hanya informasi yang bermakna yang digunakan dalam model. Sedangkan dengan mengubah semua teks menjadi huruf kecil dapat menghindari duplikasi kata dan menjaga konsistensi data sehingga dapat menyederhanakan proses tokenisasi.

2. Pelabelan Data

Kami melakukan pelabelan data pada kolom '*rating*' dimana hotel dengan *rating* 1.0 dan 2.0 diberikan label negatif dan hotel dengan *rating* 3.0, 4,0 dan 5,0 diberikan label positif. Hal ini dapat mempermudah dalam proses klasifikasi, sehingga dapat meningkatkan efesiensi proses pemodelan.

3. Menghitung Persentase Tanda Baca

Pada tahapan ini dapat memberikan informasi terhadap struktur ulasan seperti dapat mengindikasi emosi dan nada ulasan, contohnya ketika di dalam ulasan terdapat banyak tanda (!) maka dapat menunjukkan semangat atau kemarahan di dalamnya. Selain itu, persentase tanda baca dapat digunakan sebagai fitur tambahan dalam model untuk membantu klasifikasi dan prediksi. Hal ini dapat memungkinkan untuk mendeteksi apakah

ulasan tersebut ditulis oleh bot atau bertujuan untuk spam. Berikut rumus untuk menghitung persentase tanda baca:

$$\left(\frac{Jumlah\ Tanda\ Baca}{Panjang\ teks\ tanpa\ spasi}\right) \times 100$$

4. **Tokenisasi**

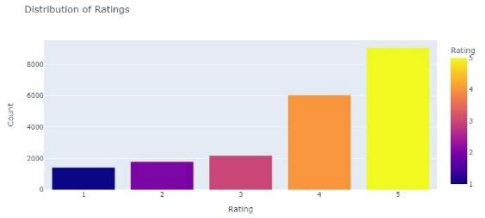
Pada tahap ini kami melakukan tokenisasi dimana memecah ulasan menjadi token-token sebelum melakukan analisis lebih lanjut.

5. **Lematisasi**

Dalam proses lematisasi ini akan mengubah setiap kata ke dalam bentuk dasar dan menghilangkan *stopwords* menggunakan *library* nltk.

C. EDA

- **Visualisasi *Distribution Of Rating***



Hasil diagram batang di atas menjelaskan penyebaran data *rating review hotel* dalam dataset. Didapatkan bahwa mayoritas *review* berisi bintang 5 yang memberikan kesan baik sekali terhadap hotel yang dikunjungi dengan jumlah rating 9.054 ulasan.

- ***Word Cloud Positive and Negative Review***

Hasil *word cloud* di atas dapat memberikan gambaran mengenai kata-kata yang sering muncul dalam ulasan. Kata-kata yang ada di dalam *word cloud* merupakan bentuk ekspresi atau kesan yang diberikan Pelanggan terhadap hotel, yang menjadi faktor penilaian atau *rating* yang diberikan Pelanggan. Dengan melihat *word cloud* tersebut pihak hotel dapat memahami preferensi Pelanggan dengan menjaga atau meningkatkan fasilitas dan layanan yang disediakan

D. Pemisahan Data

1. **Pemisahan Data Fitur dan Target**

Dalam hal ini kami memisahkan data fitur (x) dan target (y) yang mana data fitur (x) didapat dari hasil proses *preprocessing* sebelumnya yaitu kolom *'lemmatized_review', 'review_len', 'punct* sedangkan fitur (y) mengambil data pada kolom 'label'

2. **Pemisahan Data Latih dan Uji**

Dalam hal ini kami memisahkan data latih sebanyak 80% dan data uji sebanyak 20%.

E. Vektorisasi dan Reduksi Dimensi

1. **Vektorisasi menggunakan *TF-IDF Vectorizer***

Pada tahapan ini, teks akan diubah menjadi representasi numerik yang kemudian diberikan pembobotan nilai terhadap kata yang sering muncul.

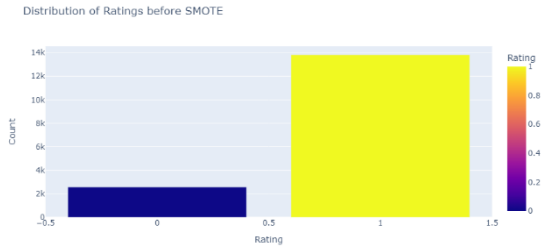
2. **Reduksi Dimensi menggunakan PCA**

Pada tahapan ini, kami mengurangi fitur dari *vector* dengan mempertahankan variasi data untuk membantu dalam meningkatkan efesiensi dan kinerja model.

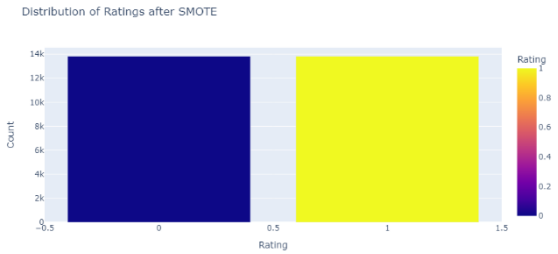
3. **Features Numeric**

Pada tahapan ini, kami menggabungkan fitur numerik dan teks yang telah diproses menjadi satu *DataFrame* untuk data pelatihan dan data uji. Langkah ini penting untuk mempersiapkan data sebelum digunakan dalam model *machine learning*, yang membutuhkan semua fitur dalam satu struktur data yang seragam.

F. Penanganan Ketidakseimbangan Data



Berdasarkan hasil EDA terdapat ketidakseimbangan kelas dalam data sehingga akan mempengaruhi proses analisis kedepannya. Oleh karena itu, kami menggunakan *over sampling SMOTE* untuk menangani ketidakseimbangan kelas tersebut.



G. Pelatihan serta Evaluasi Model *XGBoost* dan *Logistic Regression* menggunakan *Confusion Matrix*

Pada tahapan ini kami melakukan pelatihan model *XGBoost* dan *Logistic Regression* dengan data yang telah di *resample*, setelah itu kami mengevaluasi model dengan menggunakan *confusion matrix*.

IV. HASIL DAN PEMBAHASAN

Hasil dari penelitian ini berupa aplikasi streamlit yang dibangun berdasarkan model algoritma yang telah di latih. Berikut perbandingan hasil evaluasi untuk kedua algoritma:

Tabel 1. *Confusion Matrix Algoritma Logistic Regression*

	Precision	Recall	F1-score	support
Negative (0)	0.64	0.90	0.74	640
Positive (1)	0.98	0.91	0.94	3459
Accuracy			0.90	4099
Macro avg	0.81	0.90	0.84	4099
Weighted avg	0.93	0.90	0.91	4099

Tabel 1 menunjukkan kinerja model Logistic Regression dengan precision, recall, dan F1-score untuk kelas negatif (0) dan positif (1). Model memiliki akurasi keseluruhan 90% dengan kinerja terbaik pada kelas positif (precision 0.98 dan recall 0.91). Rata-rata tertimbang menunjukkan bahwa model bekerja sangat baik secara keseluruhan dengan F1-score 0.91.

Tabel 2. *Confusion Matrix Algoritma XGBoost*

	Precision	Recall	F1-score	support
Negative (0)	0.69	0.80	0.74	640
Positive (1)	0.96	0.93	0.95	3459
Accuracy			0.91	4099
Macro avg	0.83	0.87	0.84	4099
Weighted avg	0.92	0.91	0.92	4099

Tabel menunjukkan kinerja model XGBoost dengan precision, recall, dan F1-score untuk kelas negatif (0) dan positif (1). Model memiliki akurasi keseluruhan 91% dengan kinerja terbaik pada kelas positif (precision 0.96 dan recall 0.93). Rata-rata tertimbang menunjukkan bahwa model bekerja sangat baik secara keseluruhan dengan F1-score 0.92.

V. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan pada analisis sentimen ulasan hotel berbahasa Inggris menggunakan *XGBoost* dan *Logistic Regression* dapat disimpulkan bahwa pembobotan menggunakan *TF-IDF* terhadap kata yang sering muncul. Hasil dari prediksi pada metode *XGBoost* yaitu menghasilkan nilai akurasi tertinggi sebesar 91,29%. Kemudian pada evaluasi kinerja metode *Logistic Regression* menghasilkan nilai akurasi sebesar 90,29%. Oleh karena itu, dapat ditarik kesimpulan metode *XGBoost* bekerja lebih baik dalam mengklasifikasikan ulasan hotel. Namun pada penelitian ini masih terdapat kekurangan seperti adanya kalimat ulasan yang maknanya ambigu sehingga dapat meningkatkan tingkat kesalahan model. Oleh karena itu, disarankan bagi penelitian selanjutnya untuk mengekstrak aspek dari data validasi agar hasil prediksi yang diperoleh lebih akurat.

REFERENSI

[1] A. H. Hasugian, M. Fakhriza, and D. Zukhoiriyah, "Analisis Sentimen Pada Review Pengguna E-Commerce Menggunakan Algoritma Naïve Bayes," *Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD*, vol. 6, no. 1, pp. 98–107, 2023, [Online]. Available: https://ojs.trigunadharma.ac.id/index.php/jsk/index

[2] V. W. D. Thomas and F. Rumaisa, "Analisis Sentimen Ulasan Hotel Bahasa Indonesia Menggunakan Support Vector Machine dan TF-IDF," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 3, pp. 1767–1774, Jul. 2022, doi: 10.30865/mib.v6i3.4218.

[3] H. Chyntia Morama, D. E. Ratnawati, and I. Arwani, "Analisis Sentimen berbasis Aspek terhadap Ulasan Hotel Tentrem Yogyakarta menggunakan Algoritma Random Forest Classifier," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* , vol. 6, no. 4, pp. 1702–1708, 2022, [Online]. Available: http://j-ptiik.ub.ac.id

[4] R. Cahyani and P. Pandu Adikara, "Analisis Sentimen terhadap Ulasan Hotel menggunakan Boosting Weighted Extreme Learning Machine," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* , vol. 3, no. 8, pp. 2548–964, 2019, [Online]. Available: http://j-ptiik.ub.ac.id

[5] I. Santoso and L. Hulliyyatus Suadaa, "PENGUKURAN TINGKAT KEMIRIPAN DOKUMEN BERBASIS CLUSTER PADA CORPUS BESAR," *Kumpulan jurnaL Ilmu Komputer (KLIK)*, vol. 6, no. 1, pp. 71–83, 2019.

[6] Wiyanto, Wowon Priatna, and Jumi Saroh Hidayat, "IMPLEMENTASI TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF) DAN VECTOR SPACE MODEL (VSM) UNTUK PENCARIAN BERITA BAHASA INDONESIA," *Pelita Teknologi: Jurnal Ilmiah Informatika, Arsitektur dan Lingkungan*, vol. 14, no. 2, pp. 119–133, 2019, [Online]. Available: https://www.researchgate.net/publication/336982602

[7] M. Erkamim, S. Suswadi, M. Z. Subarkah, and E. Widarti, "Komparasi Algoritme Random Forest dan XGBoosting dalam Klasifikasi Performa UMKM," *Jurnal Sistem Informasi Bisnis*, vol. 13, no. 2, pp. 127–134, Oct. 2023, doi: 10.21456/vol13iss2pp127-134.

[8] S. Wahyuni Kalumbang, "PERBANDINGAN REGRESI LOGISTIK, KLASIFIKASI NAIVE BAYES, DAN RANDOM FOREST (COMPARISON THE LOGISTIC REGRESSION, NAIVE BAYES CLASSIFICATION, AND RANDOM FOREST)," *Jurnal Matematika Thales (JMT)*, vol. 03, no. 02, pp. 1–13, 2021.

[9] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," *Jurnal Sains Komputer & Informatika (J-SAKTI*, vol. 5, no. 2, pp. 697–711, 2021.