JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1551-1562
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6461

Penerapan Metode CRISP-DM dalam Klasifikasi Data Ulasan Pengunjung Destinasi Danau Toba Menggunakan Algoritma Naïve Bayes Classifier (NBC) dan Decision Tree (DT)

Yerik Afrianto Singgalen*

Fakultas Ilmu Administrasi Bisnis dan Ilmu Komunikasi, Program Studi Pariwisata, Universitas Katolik Indonesia Atma Jaya, Jakarta, Indonesia
Email: yerik.afrianto@atmajaya.ac.id
Email Penulis Korespondensi: yerik.afrianto@atmajaya.ac.id

Abstrak−Penelitian ini bertujuan untuk mengimplementasikan metode klasifikasi menggunakan algoritma Naïve Bayes Classifier (NBC) pada data teks ulasan pengunjung di Danau Toba. Metode Cross Industry Standard Process for Data Mining (CRISP-DM) terdiri beberapa tahapan sebagai berikut : tahap business understanding; tahap data understanding; tahap data preparation; tahap modeling; tahap evaluation; tahap deployment. Hasil penelitian ini menunjukkan bahwa pada tahap business understanding, konteks pembahasan menekankan pada sektor pariwisata yakni persepsi wisatawan terhadap kualitas produk dan layanan destinasi wisata Danau Toba. Pada tahap data understanding, sumber data ulasan yang digunakan berasal dari website Tripadvisor sebanyak 858 ulasan dengan klasifikasi rating sebagai berikut : 8 ulasan dengan rating sangat buruk; 22 ulasan dengan rating buruk; 81 ulasan dengan rating netral; 304 ulasan dengan rating baik; 443 ulasan dengan rating sangat baik. Pada tahap data preparation, dilakukan pembersihan data sehingga terdapat 382 data yang akan diproses dengan pembagian 30% data latih dan 70% data uji. Pada tahap modeling, dilakukan pengujian performa algoritma NBC dan DT, menggunakan operator SMOTE UPsampling dan tanpa menggunakan SMOTE UPsampling. Adapun hasil perbandingan nilai algoritma NBC dan DT menunjukkan bahwa model dengan performa terbaik ialah DT menggunakan operator SMOTE UPsampling dengan nilai akurasi (98,27%), presisi (98,83%), recall (97,71%), f-measure (98,26%), dan nilai AUC (0,982). Pada tahap evaluasi, dilakukan analisis hasil perankingan lima kata populer dalam data ulasan pengunjung Danau Toba yang memberikan penekanan pada pentingnya pelayanan prima (SDM berkualitas) dan infrastruktur pendukung (sarana dan prasarana pariwisata). Pada tahap deployment, dibutuhkan keseimbangan dalam pengembangan atraksi, aksesiblitas, akomodasi dan amentias pendukung pariwisata agar dapat memantik minat berkunjung dan motivasi untuk berkunjung kembali ke Danau Toba.

Kata Kunci: Analisis Sentimen; CRISP-DM; Danau Toba; Klasifikasi; NBC; DT

Abstract−This study aims to implement a classification method using the Nave Bayes Classifier (NBC) algorithm on Lake Toba visitor review text data. The Cross Industry Standard Process for Data Mining (CRISP-DM) methodology comprises the following stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The findings of this study indicate that during the phase of business comprehension, the context of the discussion focuses on the tourism sector, specifically tourist perceptions of the quality of products and services at Lake Toba tourist destinations. At the data comprehension stage, the source of review data used was the Tripadvisor website, which contained as many as 858 reviews with the following rating classification: 8 reviews with abysmal ratings; 22 reviews with poor ratings; 81 reviews with neutral ratings; 304 reviews with good ratings; 443 reviews with excellent ratings. Data cleansing is performed at the data preparation stage so that 382 data are processed by dividing training data by 70 percent and test data by 30 percent. During the modeling phase, the performance of the NBC and DT algorithms was evaluated using and without SMOTE UPsampling operators. The comparison of NBC and DT algorithm values indicates that the model with the best performance is DT using SMOTE UPsampling operators with accuracy values (98.27 percent), precision values (98.83 percent), recall values (97.71 percent), f-measure values (98.26 percent), and AUC values (98.27 percent) (0.982). At the evaluation stage, the importance of excellent service (Quality Human Resources) and supporting infrastructure was highlighted by analyzing the results of ranking the five most frequently used terms in Lake Toba visitor review data (tourism facilities and infrastructure). At the deployment stage, it is necessary to balance the development of attractions, accessibility, lodging, and tourism-supporting amenities to generate visiting intention and revisit motivation to Lake Toba.

Keywords: Sentiment Analysis; CRISP-DM; Toba Lake; Classification; NBC; DT

1. PENDAHULUAN

Analisis sentimen merupakan salah satu pendekatan yang efektif digunakan dalam mengidentifikasi preferensi dan kepuasan wisatawan terhadap produk dan layanan di berbagai destinasi wisata. Meturan et al. menunjukkan bahwa ketersediaan sarana dan prasarana di destinasi wisata menjadi salah satu faktor yang menentukan kepuasan pengunjung [1]. Disisi lain, Azzahra dan Wibowo menunjukkan bahwa atraksi, aksesibilitas, akomodasi dan amenitas di suatu destinasi wisata dapat memantik persepsi wisatawan sekaligus menunjukkan sentimen wisatawan selama berwisata [2]. Hal ini menunjukkan bahwa sentimen wisatawan yang diklasifikasi berdasarkan sentimen positif dan sentimen negatif dapat digunakan sebagai evaluasi manajemen destinasi wisata. Adapun, Riadi et al. menegaskan bahwa kepuasan wisatawan memiliki pengaruh yang signifikan terhadap minat berkunjung kembali [3]. Dengan demikian dapat diketahui bahwa pengelolaan destinasi wisata yang optimal dapat mendukung kepuasan wisatawan serta memantik minat berkunjung Kembali. Mempertimbangkan hal tersebut maka perlu dilakukan kajian tentang analisis sentimen wisatawan yang berkunjung ke destinasi wisata Danau Toba.

Submitted: **26/06/2023**; Accepted: **31/07/2023**; Published: **31/07/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1551-1562
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6461

Salah satu metode yang digunakan dalam analisis sentimen ialah Cross Industry Standard Process for Data Mining (CRISP-DM) yang terdiri dari tahap business understanding, tahap data understanding, tahap data preparation, tahap modeling, tahap evaluation, dan tahap deployment. Christian dan Qi menggunakan metode CRISP-DM dalam menganalisis sengmen pasar menggunakan algoritma k-Means sebagai model [4]. Disisi lain, Munawwaroh dan Primandani menggunakan metode CRISP-DM dalam memprediksi Lingkar Lengan Atas (LILA) Ibu Hamil berpotensi kurang gisi melalui model Decision Tree (DT) [5]. Hal ini menunjukkan bahwa penggunaan CRISP-DM sebagai metode yang fokus pada pemahaman proses bisnis dari suatu organisasi secara kelembagaan, serta memperitmbangkan karakteristik data yang digunakan dalam proses modeling. Dengan demikian, luaran dari proses pengolahan data dapat menghasilkan interpretasi terhadap masalah yang lebih spesifik serta merekomendasikan solusi yang relevan dengan kebutuhan institusi untuk mengoptimalkan performa organisasi. Mempertimbangkan hal tersebut maka penelitian ini fokus pada konteks kepariwisataan yakni manajemen destinasi wisata Danau Toba dan persepsi wisatawan yang terejawantahkan dalam data ulasan di website Tripadvisor. Adapun, algoritma yang digunakan ialah Naïve Bayes Classifier (NBC) dan Decission Tree (DT) dalam penerapan metode klasifikasi sentimen.

Perilaku wisatawan dalam merencanakan perjalanan wisata ialah proses penelusuran informasi melalui media sosial hingga website yang merekomendasikan destinasi wisata menarik dengan fasilitas lengkap. Harnawi mengemukakan bahwa website Tripadvisor menjadi salah satu platform digital yang paling banyak digunakan sebagai rujukan untuk merencanakan perjalanan wisata [6]. Disisi lain, Rita et al. menunjukkan bahwa fitur online review dan sistem rating yang diterapkan dalam website Tripadvisor berdampak pada penilaian pengguna sistem serta memengaruhi persepsi viewer [7]. Hal ini menunjukkan bahwa website Tripadvisor dapat digunakan sebagai sumber data teks untuk mengklasifikasi sentimen pengunjung, serta menganalisis secara komprehensif aspek-aspek yang memengaruhi perilaku wisatawan. Dengan demikian dapat diketahui bahwa website Tripadvisor dapat digunakan sebagai sumber data teks untuk diproses menggunakan metode CRISP-DM melalui algoritma NBC dan DT.

Dalam konteks kepariwisataan, brand destinasi dan layanan akomodasi di website Triapdvisor berperan penting dalam meningkatkan performa bisnis. Dewi et al. menunjukkan bahwa brand image dari penyedia layanan akomodasi di situs Tripadvisor dapat memengaruhi daya beli serta memantik perilaku konsumen dalam menyebarkan informasi di lingkungan sosial masing-masing [8]. Disisi lain, Lynn et al. menunjukkan bahwa narasi terkait dengan fasilitas dan layanan akomodasi di website Tripadvisor memiliki pengaruh yang signifikan terhadap calon konsumen, sehingga brand yang negatif cenderung memengaruhi motivasi dan minat beli pengguna website Tripadvisor [9]. Adapun, Minkwitz menegaskan bahwa Tripadvisor dapat digunakan sebagai sumber data dalam merencanakan program pengembangan destinasi wisata lokal [10] Hal ini menunjukkan bahwa masing-masing penyedia layanan destinasi wisata, layanan akomodasi dan amenitas, serta layanan transportasi perlu mengoptimalkan produk dan layanan sehingga memberikan kesan yang positif bagi konsumen. Dengan demikian, ulasan dalam bentuk narasi yang positif di website Tripadvisor dapat memantik motivasi berkunjung atau minat beli konsumen terhadap produk maupun layanan yang ditawarkan.

Dalam konteks Indonesia, kepariwisataan berperan penting dalam meningkatkan devisa negara, mendorong pertumbuhan ekonomi, serta mewujudkan kesejahteraan sosial dan keberlanjutan ekologi. Selain itu, intensitas penggunaan media digital di Indonesia mengalami peningkatan yang signifikan. Pramudita et al. mengemukakan bahwa pengguna internet di Indonesia meningkat seiring dengan optimalisasi infrastruktur teknologi dan kebutuhan penggunaan informasi digital [11]. Disisi lain, Santoso et al. menunjukkan adanya perubahan perilaku masyarakat seiring dengan perkembangan teknologi [12]. Dalam konteks kepariwisataan, perilaku penggunaan teknologi digital memungkinkan terjadinya intensitas penggunaan website Tripadvisor sebagai sumber data dan informasi untuk merencanakan perjalanan wisata, serta rujukan penggunaan layanan akomodasi dan transportasi. Goldsmith menunjukkan bahwa komentar negatif terhadap layanan akomodasi di website Tripadvisor memengaruhi brand image layanan akomodasi dan menimbulkan kesan tidak professional yang berdampak minat beli calon konsumen [13]. Hal ini menunjukkan bahwa pemangku kepentingan perlu meningkatkan brand image bisnis pendukung pariwisata, serta memanfaatkan media digital sebagai media pemasaran sehingga meningkatkan minat berkunjung. Dengan demikian dapat diketahui adanya peluang untuk mengoptimalkan pariwisata Indonesia melalui teknologi informasi.

Analisis sentimen wisatawan perlu dilakukan secara berkala untuk mengidentifikasi perubahan preferensi wisatawan terkait dengan produk dan layanan yang berhubungan dengan atraksi, aksesibilitas, akomodasi dan amenitas. Khoffifah et al. menggunaan algoritma NBC dalam menganalisis sentimen wisatawan terhadap produk dan layanan di Kabupaten Karawang [14]. Selanjutnya, Pati dan Umar melalukan analisis sentimen menggunakan algoritma NBC dan k-Nearest Neighbor (kNN) terhadap destinasi wisata Danau Weekuri [15]. Hal ini menunjukkan bahwa analisis sentimen di bidang pariwisata bermanfaat sebagai fungsi pengendalian kualitas produk dan layanan. Disisi lain, Christanto et al. menunjukkan bahwa perbandingan algoritma NBC, Decision Tree (DT) dan Xgboost yang digunakan dalam klasifikasi sentimen wisatwan berperan penting dalam proses evaluasi model dengan performa terbaik [16]. Ginantra et al. melakukan analisis sentimen dengan membandingkan perfroma algoritma NBC, DT dan k-NN untuk mengevaluasi model terbaik dalam klasifikasi sentimen tamu hotel di Ubud. Hal ini menunjukkan bahwa kajian tentang analisis sentimen berpeluang memberikan kontribusi secara empiris dan teoretis di bidang manajamen produk dan layanan pariwisata. Dengan demikian, penelitian ini

Submitted: **26/06/2023**; Accepted: **31/07/2023**; Published: **31/07/2023**

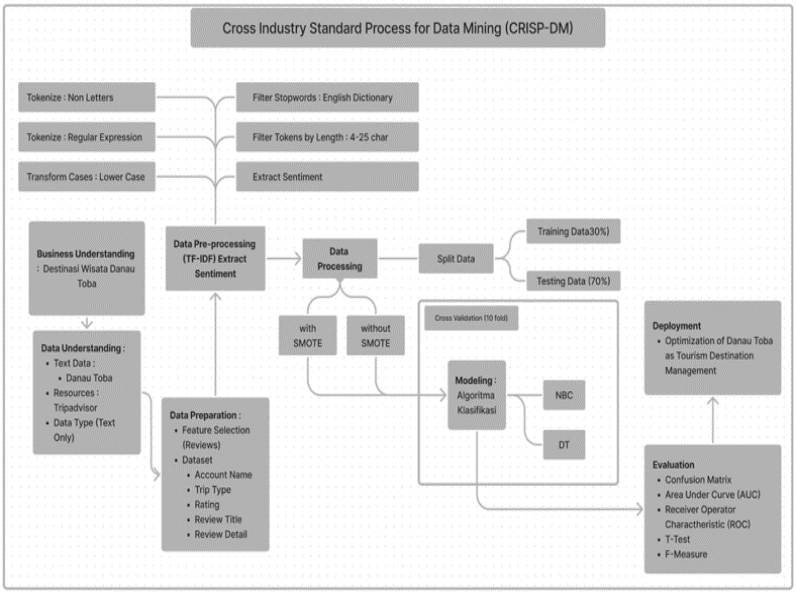
JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1551-1562
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6461

bertujuan menganalisis sentimen pengunjung destinasi wisata Danau Toba sebagai salah satu destinasi prioritas nasional di Indonesia, untuk mengidentifikasi preferensi wisatawan serta menganalisis secara komprehensif peluang dan tantangan manajemen destinasi wisata Danau Toba berdasarkan data ulasan pengunjung di website Tripadvisor.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian Berdasarkan Cross Industry Standard Process for Data Mining (CRISP-DM)

Penelitian ini mengadopsi metode Cross Industry Standard Process for Data Mining (CRISP-DM) yang terdiri dari tahap business understanding, tahap data understanding, tahap data preparation, tahap modeling, tahap evaluation, dan tahap deployment. Pertimbangan menggunakan metode CRISP-DM ialah sebagai berikut : pertama, metode CRISP-DM secara spesifik menekankan pada konteks data dan tujuan pengolahan data sehingga fokus utama pembahasan terkait kualitas produk dan layanan yang berhubungan dengan manajemen destinasi wisata, didukung oleh data yang relevan dan koheren; kedua, sumber data dan proses persiapan data dalam metode CRISP-DM sangat fleksibel dan terukur, serta memudahkan data analis dalam memilih sumber yang valid dan kredibel sebelum proses pemodelan; ketiga, pemodelan dilakukan berdasarkan performa algoritma terbaik yang relevan dengan tujuan pengolahan data; keempat, proses evaluasi dan deployment dapat disesuaikan dengan kebutuhan masing-masing institusi atau organisasi yang menggunakan hasil interpretasi terhadap data sentimen. Meskipun demikian, implementasi metode CRISP-DM dalam kajian tentang analisis sentimen destinasi pariwisata di Indonesia, masih sangat terbatas. Beberapa penelitian terdahulu memberikan penekanan langsung pada algoritma yang digunakan. Arifiyanti et al. melakukan analisis sentimen ulasan pengunjung destinasi wisata Gunung Bromo berdasarkan data ulasan yang bersumber dari website Tripadvisor [17]. Disisi lain, Somantri dan Dairoh mengimplementasikan teknik penambangan data (data mining) dan mengolah data teks dengan metode klasifikasi menggunakan algoritma NBC dan DT untuk menganalisis sentimen wisatawan di kota Tegal [18]. Hal ini menunjukkan bahwa metode CRISP-DM tidak selalu digunakan dalam proses klasifikasi sentimen. Dengan demikian, penelitian ini menggunakan CRISP-DM dalam konteks analisis sentimen untuk menghasilkan informasi yang valid dan kredibel sebagaimana tahapan-tahapan dalam metode CRISP-DM, sebagiamana gambar berikut.

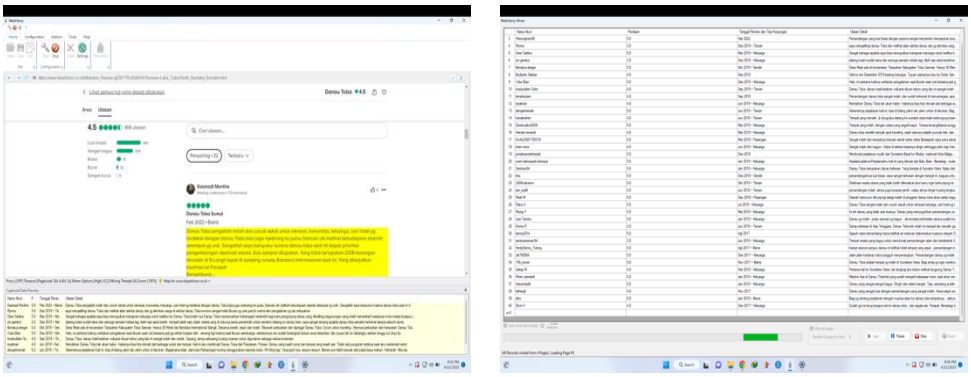


Gambar 1. Tahapan dalam Metode CRISP-DM

Gambar 1 merupakan implementasi metode CRISP-DM berdasarkan setiap tahapan dalam proses analisis sentimen wisatawan destinasi wisata Danau Toba, Indonesia. Pada tahap business understanding, penetapan konteks pembahasan fokus pada isu manajemen destinasi wisata yang dapat ditinjau berdasarkan persepsi wisatawan melalui data teks berupa ulasan pengunjung. Tahap data understanding, merupakan proses identifikasi dan analisis sumber data yakni website Tripadvisor. Berdasarkan hasil identifikasi dan analisis alur reviews di website Tripadvisor, dapat diketahui bahwa sebelum ulasan dipublikasikan, proses verifikasi dan validasi dilakukan oleh pengelola website Tripadvisor agar informasi yang terpublikasi di kolom reviews, dapat dipertanggungjawabkan oleh masing-masing user. Dengan demikian, data teks di kolom online reviews serta rating yang diberikan oleh user, dapat digunakan sebagai sumber data untuk diproses lebih lanjut melalui klasifikasi sentimen wisatawan. Pada tahap data preparation, proses scraping data teks dilakukan menggunakan aplikasi webharvy, dengan proses seleksi data berdasarkan nama akun, rating, tanggal ulasan dan tipe kunjungan, serta deskripsi ulasan, sebagaimana gambar berikut.

Submitted: **26/06/2023**; Accepted: **31/07/2023**; Published: **31/07/2023**

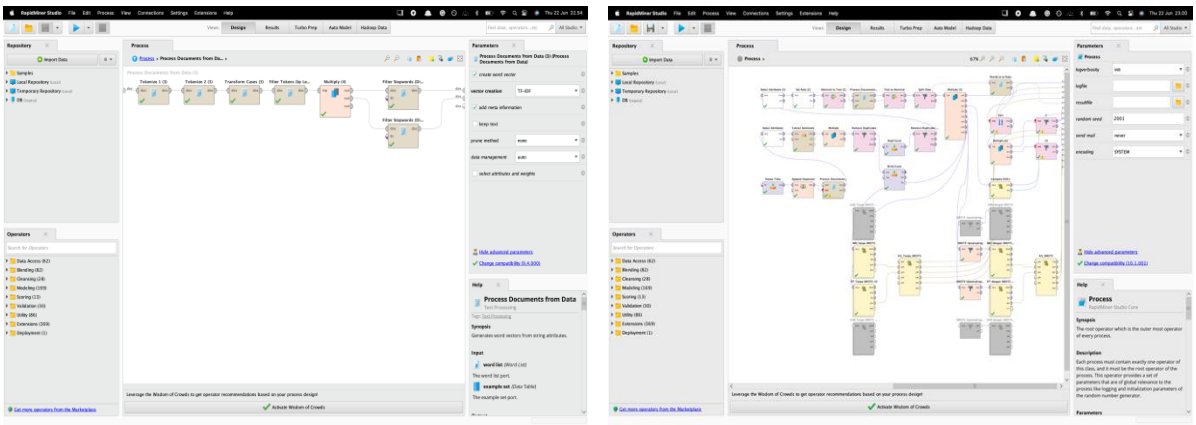
JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1551-1562
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6461



Gambar 2. Proses Scraping Data Ulasan dari Website Tripadvisor Menggunakan Webharvy

Gambar 2 merupakan proses scraping data ulasan dari website Tripadvisor menggunakan aplikasi Webharvy dengan klasifikasi data sesuai nama akun, tanggal ulasan dan tipe kunjungan, serta deskripsi ulasan. Berdasarkan hasil identifikasi sumber data dari website Tripadvisor melalui link (https://www.tripadvisor.co.id/Attraction_Review-g2301775-d338410-Reviews-Lake_Toba-North_Sumatra_Sumatra.html) terdapat 858 ulasan yang terpublikasi dengan klasifikasi rating sebagai berikut : 8 ulasan dengan rating sangat buruk; 22 ulasan dengan rating buruk; 81 ulasan dengan rating netral; 304 ulasan dengan rating baik; 443 ulasan dengan rating sangat baik. Secara spesifik, tidak ditampilkan klasifikasi rating berdasarkan tipe kunjungan (bisnis, pasangan, keluarga, dengan teman, sendiri), serta tidak diklasifikasikan berdasarkan rating, tipe kunjungan, tanggal dan tahun berkunjung. Hal ini menunjukkan bahwa website Tripadvisor terbatas menampilkan visualisasi data rating pengunjung yang terpublikasi berdasarkan tipe kunjungan, tanggal dan tahun berkunjung tanpa informasi yang spesifik terkait tujuan visualisasi dan publikasi data. Mempertimbangkan hal tersebut, maka data teks dari Tripadvisor dapat diproses menggunakan metode CRISP-DM untuk menghasilkan informasi yang bermanfaat bagi pengelola destinasi wisata Danau Toba dalam meningkatkan kualitas produk dan layanan sebagaimana preferensi wisatawan.

Data yang telah dikumpulkan akan disiapkan terlebih dahulu melalui proses data pre-processing pada aplikasi Rapidminer. Plugin yang digunakan untuk data pre-processing ialah sebagai berikut : pertama, tokenize dengan non-letters mode untuk menghilangkan angka dalam ulasan, sehingga hanya teks yang akan diproses; kedua, tokenize dengan regular expression mode untuk menghilangkan simbol-simbol dan angka; ketiga, transform cases untuk mengubah masing-masing kata menjadi lowercase; keempat, filter tokens by length dengan minimal karakter 4 dan maksimal 25; filter stopwords menggunakan kamus bahasa inggris dan bahasa Indonesia. Selanjutnya dilakukan proses pembersihan data duplikan menggunakan remove duplicate plugin pada aplikasi Rapidminer. Adapun, proses pre-processing dan processing data dapat dilihat pada gambar berikut.



Gambar 3. Pre-Processing dan Processing Data Menggunakan RapidMiner

Gambar 3 merupakan alur pre-processing dan processing data menggunakan aplikasi Rapidminer. Berdasarkan hasil scraping data, dapat diketahui bahwa terdapat 424 data teks ulasan wisatawan yang dapat dilanjutkan ke tahap pre-processing. Selanjutnya, terdapat 382 data ulasan hasil pre-processing yang dapat dilanjutkan ke tahap data processing menggunakan algoritma NBC dan DT. Performa algoritma terbaik dapat digunakan sebagai model yang relevan dalam pengolahan data ulasan pengunjung Danau Toba. Pada tahap evaluasi, hasil confusion matrix masing-masing algoritma yang menampilkan informasi tentang nilai accuracy, recall, precision, Area Under Curve (AUC), f-measure dapat diperbandingkan dan dianalisis secara komprehensif terkait konteks dataset. Selanjutnya, hasil ekstrak sentimen dalam bentuk 10 dan 50 kata populer divisualisasikan dalam bentuk wordcloud sebagaimana gambar berikut.

Submitted: **26/06/2023**; Accepted: **31/07/2023**; Published: **31/07/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1551-1562
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6461

Gambar 4. Wordcloud 10 and 50 Kata Popular

Gambar 4 merupakan visualisasi 10 dan 50 kata populer dalam data ulasan pengunjung destinasi wisata Danau Toba. Berdasarkan lima kata populer terdapat kata-kata sebagai berikut : danau (329), toba (254), pulau (156), samosir (154), tempat (143). Hal ini menunjukkan bahwa atraksi menjadi aspek penting dalam persepsi wisatawan, sehingga perlu dikelola dengan optimal. Selanjutnya, berdasarkan 50 kata populer, terdapat beberapa kata yang mencerminkan persepsi terhadap aksesbilitas, akomodasi dan amenitas yang terjewantahkan dalam kata-kata berikut : jalan (63); kapal (53); mobil (34); jauh (39) ; dan hotel (69). Secara keseluruhan kata-kata yang sering muncul, tergolong sentimen positif. Hal ini menunjukkan bahwa data teks dalam visualisasi wordcloud dapat digunakan sebagai rujukan untuk mengevaluasi kualitas produk dan layanan di destinasi wisata Danau Toba secara berkala. Pada tahap deployment, luaran dari analisis sentimen dapat digunakan sebagai pertimbangan dalam pengambilan keputusan, serta penetapan program prioritas untuk mengoptimalkan tatakelola destinasi Danau Toba.

2.2 Algoritma Naïve Bayes Classifier

Naive Bayes Classifier (NBC) memiliki keunggulan tersendiri karena data diklasifikasikan dengan probabilitas sederhana, yang menerapkan teorema Bayes dengan asumsi independensi yang tinggi. [19]. Studi ini didasarkan pada jumlah dataset yang digunakan dalam metode dengan kemampuan klasifikasi yang cepat dan akurat. Pengklasifikasi Naive Bayesian hanya membutuhkan data pelatihan dalam jumlah yang relatif kecil untuk menentukan estimasi parameter yang diperlukan untuk proses klasifikasi. Pada tahap klasifikasi, nilai kelas ditentukan dari data berdasarkan suku yang terjadi dengan menggunakan persamaan berikut.

$$P(X_k|Y) = \frac{P(Y|X_k)}{\sum iP(Y|X_i)} \quad (1)$$

Dimana, keadaan posterior (Probabilitas X_k di dalam Y) dapat dihitung dari keadaan prior (Probabilitas Y di dalam X_k) dibagi dengan jumlah dari semua probabilitas Y di dalam semua X_i. Dalam konteks penelitian ini, data teks yang diperoleh dari website Tripadvisor diklasifikasi menggunakan persamaan berikut.

$$\text{P (v1|C=c)} = \frac{CountTerms(v1,docsv(c))}{AllTerms(docs(c))} \quad (2)$$

Dimana v1 merupakan salah satu suku kata yang muncul dalam ulasan pengguna website Tripadvisor terhadap kualitas produk dan layanan di Destinasi Wisata Candi Borobudur. Sedangkan, $CountTerms(v1, docsv(c))$ merujuk pada jumlah kemunculan suatu kata berlabel c ("positif" atau "negatif"). Adapun, $AllTerms(docs(c))$ merujuk pada jumlah semua kata berlabel c yang ada pada dataset. Untuk menghindari adanya nilai nol pada probabilitas maka diimplementasikan laplace smoothing, untuk mengurangi probabilitas dari hasil yang terobservasi, dan jua meningkatkan probabilitas hasil yang belum terobservasi. Dengan demikian, persamaan yang digunakan ialah sebagai berikut :

$$\text{P (v1|C=c)} = \frac{CountTerms(v1,docsv(c))+1}{AllTerms(docs(c))+|V|} \quad (3)$$

Dimana |V| merujuk pada jumlah semua kata dalam data ulasan yang ada di dataset. Dengan demikian, proses klasifikasi data ulasan akan menunjukkan kata dengan nilai tertinggi sebagai representasi perhatian pengulas terhadap produk dan layanan yang diperoleh.

2.3 Algoritma Decision Tree

Decision Tree (DT) merupakan model prediksi yang menggunakan struktur pohon untuk mencari dan membuat keputusan, serta memecahkan masalah dengan mempertimbangkan berbagai faktor di dalam lingkup masalah tersebut [20]. Decision Tree memiliki beberapa algoritma salah satunya Interative Dychotomizer version (ID3), yaitu model klasifikasi yang berupa pohon keputusan secara top-down dengan cara kerja mengevaluasi semua atribut menggunakan suatu ukuran statistik berupa information gain untuk mengukur efektifitas suatu atribut dalam

Submitted: **26/06/2023**; Accepted: **31/07/2023**; Published: **31/07/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1551-1562
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6461

mengklasifikasi sample data. Dalam algoritma ini, dibutuhkan nilai entropy dan gain, dimana entropy merupakan parameter untuk mengukur jumlah keberagaman atau keberadaan dalam sebuah himpunan data, sedangkan gain merupakan perolehan informasi sebagai ukuran efektifitas suatu atribut. Berikut adalah persamaan untuk mendapatkan nilai entropy dan gain.

$$Entropy\ (S) = \sum_{i=1}^{n} -p\ x\ log_2\ pi \quad (4)$$

$$Gain\ (S, A) = S - \sum_{i=1}^{n} \frac{|Si|}{|S|}\ x\ Si \quad (5)$$

Dimana S merupaan nilai Entropy, pi jumlah yang memiliki nilai positif atau negatif pada kumpulan data untuk sifat tertentu. Disisi lain, Gain (S,A) adalah hasil informasi yang berasal dari luaran data yang dikelompokan sesuai dengan atribut A. Selanjutnya, Si adalah subset dari nilai entropy yang mempunyai nilai i. Adapun, S adalah subset dari nilai Entropy.

2.4 Term Frequency-Inverse Document Frequency (TF-IDF)

Term Frequency-Inverse Document Frequency (TF-IDF) adalah salah satu teknik pemrosesan bahasa alami yang umum digunakan dalam analisis teks, termasuk analisis sentimen. Vitandi et al. menjelaskan bahwa TF-IDF digunakan untuk mengukur seberapa penting suatu kata dalam dokumen atau korpus melalui tahapan dalam proses sebagai berikut : tokenisasi; pre-processing; menghitung Term Frequency (TF); menghitung Inverse Document Frequency (IDF); mengalikan TF dan IDF. Pada tahap tokenisasi, dokumen atau teks yang akan dianalisis dibagi menjadi token, yakni unit terkecil yang bisa diolah oleh program. Pada tahap ini, tanda baca, angka, dan kata penghubung biasanya dihilangkan. Selanjutnya, pada tahap preprocessing, token-token yang telah dihasilkan pada tahap sebelumnya diproses lebih lanjut, seperti menghilangkan kata-kata yang tidak penting (stop words), mengubah kata menjadi bentuk dasar (stemming), dan melakukan normalisasi seperti mengubah kata yang ditulis dengan huruf besar menjadi huruf kecil. Selanjutnya, pada tahap menghitung term frequency (TF), frekuensi kemunculan setiap token dalam dokumen dihitung. Hal ini dilakukan untuk mengetahui seberapa sering sebuah kata muncul dalam dokumen. Kemudian pada tahap menghitung inverse document frequency (IDF), bobot diberikan pada setiap kata berdasarkan seberapa umum kata tersebut dalam korpus teks. Kata-kata yang jarang muncul pada korpus akan memiliki bobot yang lebih tinggi daripada kata-kata yang sering muncul. Adapun, tahap akhirnya ialah mengalikan TF dengan IDF, nilai TF dan IDF dikalikan untuk menghasilkan skor TF-IDF yang akan digunakan sebagai dasar untuk menganalisis sentimen [21].

Pada kerangka kerja CRISP-DM proses TF-IDF dilakukan untuk memahami data dan menyipakan data (tahap data understanding dan data preparation), proses seleksi diperlukan untuk membersihkan dan merapikan data ulasan yang telah dikumpulkan. Selanjutnya, pembobotan kata diperlukan untuk memperoleh informasi tentang jumlah kata yang paling sering muncul dalam data ulasan. Pembobotan kata merupakan proses pemberian nilai pada setiap kata yang telah melewati tahap Pre-Processing. Penelitian ini menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF) dengan persamaan (6) dan (7) berikut :

$$IDF(w) = log(\frac{N}{DF(w)}) \quad (6)$$

$$W_{ij} = TF_{ij} x\ log(^{D}/_{DF_j}) \quad (7)$$

Kosasih dan Alberto menunjukkan bahwa roses pembobotan kata dapat dilakukan dengan beberapa tahapan sebagai berikut : pertama, menghitung jumlah Term Frequency (FT) tiap kata, dimana kalimat yang telah dipisah menjadi kata akan diberi nilai dan setiap kata yang muncul akan diberi nilai 1; kedua, menghitung jumlah Document Frequency (DF) tiap kata dengan cara menjumlahkan nilai TF pada tiap kata; ketiga, menghitung jumlah Inverse Document Frequency (IDF) yang ditunjukkan pada persamaan (1); keempat, menghitung bobot (Weight) pada tiap kata yang diperoleh dari hasil perkalian nilai TF dengan IDF sebagaimana persamaan (7) [22]. Ardiansyah et al. menjelaskan bahwa TF-IDF merupakan metode yang berguna untuk mengekstrak fitur-fitur penting dari teks dalam analisis sentimen [23]. Dalam analisis sentimen, TF-IDF dapat membantu mengidentifikasi kata-kata atau frasa-frasa yang paling sering muncul pada teks dengan sentimen positif atau negatif, dan dapat digunakan sebagai fitur pada model klasifikasi [24]. Dengan menggunakan metode ini, kita dapat memperoleh informasi yang lebih relevan dan bermanfaat dari data teks, sehingga dapat meningkatkan kinerja model klasifikasi dalam memprediksi sentimen dari teks baru.

2.5 Synthetic Minority Oversampling Technique (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) adalah teknik resampling data yang umum digunakan dalam machine learning, khususnya untuk menangani masalah ketidakseimbangan kelas [25]. SMOTE menghasilkan sampel baru yang mirip dengan data minoritas dengan cara membuat sampel sintetis baru dari data yang sudah ada. Barro et al. berpendapat bahwa ketidakseimbangan data akan terjadi apabila jumlah objek di suatu kelas data memiliki kuantitas yang lebih tinggi dibandingkan dengan kelas lain, dimana kelas data yang objeknya lebih banyak disebut kelas mayor sedangkan yang lain disebut minor [26]. Disisi lain, Kurniawati menekankan

Submitted: **26/06/2023**; Accepted: **31/07/2023**; Published: **31/07/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1551-1562
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6461

bahwa pengolahan algoritma yang tidak mempertimbangkan ketidakseimbangan data cenderung menitikberatkan kelas mayor dan bukan kelas minor, oleh sebab itu diperlukan teknik SMOTE yang menggunakan metode oversampling untuk memperbanyak pengamatan secara acak dengan menambah jumlah data kelas minor (data buatan) agar setara dengan kelas mayor [27]. Adapun, data buatan atau sintesis tersebut dibuat berdasarkan k-tetangga terdekat (k-Nearest Neighbor). Pembangkit data buatan yang berskala numerik diukur jarak kedekatannya dengan jarak euclidean sedangkan data kategorik berdasarkan kelas minor yang peubahnya berskala kategorik, dilakukan dengan rumus Value Difference Metric (VDM) yaitu :

$$\Delta(x,y) = w_x w_y \sum_{i=1}^{N} \delta(x_i y_i)^r \quad (8)$$

Persamaan (7) merupakan proses untuk membangkitkan data numerik. Dimana $\Delta(x,y)$ adalah jarak antara amatan x dengan y, sementara $w_x w_y$ merupakan bobot amatan (dapat diabaikan), N merupakan banyaknya pebuah penjelas, r bernilai 1 (jarak manhattan) atau 2 (jarak euclidean), serta $\delta(x_i y_i)^r$ jarak antar kategori. Adapun, proses pembangkit data buatan (sintesis) untuk data numerik dilakukan dengan menghitung perbedaan antar vektor utama dengan k-tetangga terdekatnya, kalikan perbedaan dengan angka yang diacak diantara 0 dan 1, kemudian tambahkan perbedaan tersebut ke dalam nilai utama pada vektor utama asal sehingga diperoleh vektor utama yang baru. Selanjutnya, pembangkit data kategorik dapat dilakukan melalui persamaan (9) sebagai berikut.

$$\delta(V_1 V_2) = \sum_{i=1}^{n} \left| \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} \right|^K \quad (9)$$

Dimana, $\delta(V_1 V_2)$ merupakan jarak antara nilai V_1 dan V_2 sedangkan C_{1i} merupakan banyaknya V_1 yang termasuk kelas I, dan C_{2i} merupakan banyaknya V_2 yang termasuk kelas I. Sementara itu, i merupakan banyaknya kelas, C_1 banyaknya nilai 1 terjadi, C_2 banyaknya nilai 2 terjadi, n merupakan banyaknya kategori, dan k merupakan konstansa. Proses pembangkitan data buatan (sintesis) untuk data kategori dilakukan dengan memilih mayoritas antara vektor utama yang dipertimbangkan dengan k-tetangga terdekatnya untuk nilai nominal, jika nilai sama maka akan dipilih secara acak. Dengan demikian, SMOTE dapat membantu meningkatkan kinerja model klasifikasi dalam memprediksi sentimen dari teks baru dengan kelas minoritas yang lebih seimbang.

2.6 Evaluasi Klasifikasi (Confusion Matrix)

Pada tahap evaluasi, setiap algoritma akan dievaluasi berdasarkan nilai akurasi, presisi, recall, dan nilai f-measure. Evaluasi klasifikasi didasarkan pada pengujian pada objek yang benar dan objek yang salah. Validasi digunakan untuk menentukan jenis model yang terbaik melalui confusion matrix sebagai informasi mengenai hasil klasifikasi actual yang dapat diprediksi oleh suatu sistem melalui nilai akurasi, presisi, dan recall, melalui persamaan berikut.

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \quad (10)$$

$$Presisi/Specificity = \frac{TP}{TP+FP} \quad (11)$$

$$Recall/Sensitivity = \frac{TP}{TP+FN} \quad (12)$$

$$f-measure = \frac{2x(Presisi\ x\ recall)}{presisi+recall} \quad (13)$$

Ginantra et al., berpendapat bahwa confusion matrix merupakan gambaran akan akurasi, presisi, recall dari proses klasifikasi data. Akurasi adalah ketepatan sistem dalam melakukan proses klasifikasi dengan benar; presisi atau sensitivity adalah rasio jumlah dokumen yang relevan dengan total jumlah dokumen yang ditemukan pada sistem klasifiaksi; recall atau spesificity adalah rasio jumlah dokumen yang ditemukan kembali oleh sistem klasifikasi dengan total jumlah dokumen yang relevan; f-measure adalah metrik evaluasi yang populer untuk menangani masalah imbalance class dengan mengombinasi recall/sensivitas dan presisi sehingga menghasilkan metrik yang efektif untuk mencari kembali informasi dalam himpunan yang tidak seimbang [28]. Dengan demikian, evaluasi performa algoritma terbaik dapat direkomendasikan sebagai model yang relevan dengan dataset untuk memperoleh luaran analisis sentimen yang tergolong fit atau sesuai.

3. HASIL DAN PEMBAHASAN

Destinasi wisata Danau Toba merupakan salah satu destinasi wisata super priotias di Indonesia. Pardosi et al. mengemukakan bahwa pengembangan produk dan layanan destinasi wisata Danau Toba didukung oleh Sumber Daya Manusia (SDM) yang memiliki pengetahuan dan keterampilan di bidang hospitality, oleh sebab itu dukungan berbagai institusi pendidikan formal dan informal dalam pengembangan SDM menjadi faktor penentu keberhasilan dan keberlanjutan pariwisata [29]. Disisi lain, Wulandari et al. menunjukkan bahwa akselerasi pengembangan destinasi wisata Danau Toba juga dikembangkan melalui prinsip-prinsip good governance pada badan otorita Danau Toba [30]. Hal ini menunjukkan bahwa optimalisasi manajemen destinasi wisata Danau Toba tidak terlepas dari berbagai program pengembangan SDM dan optimalisasi fungsi organisasi atau lembaga yang memiliki

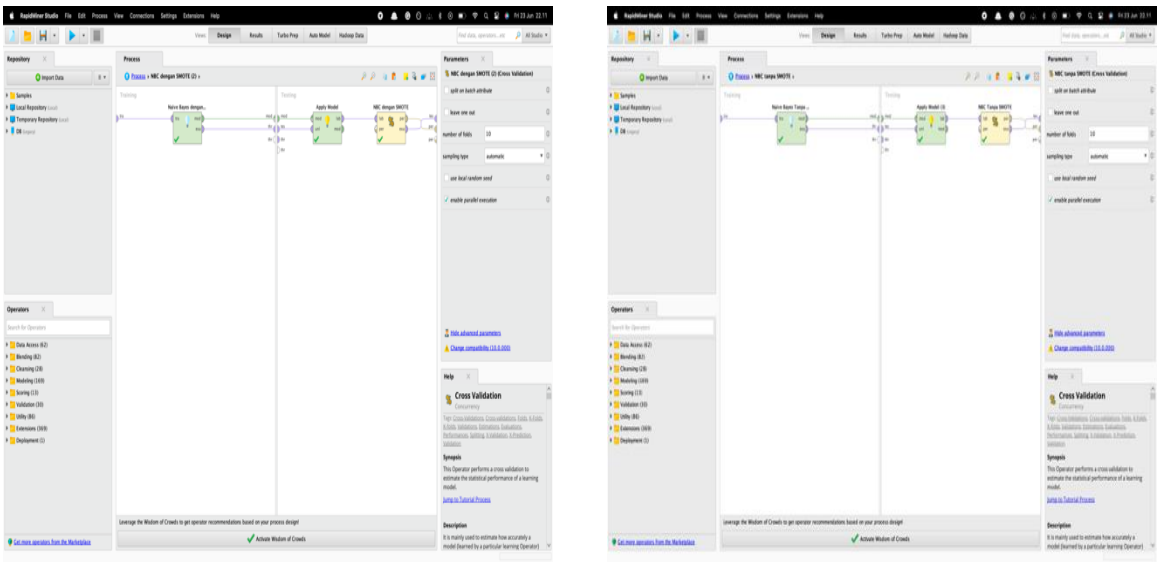
Submitted: **26/06/2023**; Accepted: **31/07/2023**; Published: **31/07/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1551-1562
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6461

kewenangan dalam hal pengembangan wilayah sekitara Danau Toba. Dengan demikian, dapat diketahui bahwa optimalisasi manajemen destinasi Danau Toba tidak terlepas dari kolaborasi para pemangku kepentingan.

Penetapan destinasi wisata Danau Toba sebagai destinasi wisata super prioritas dapat memantik pertumbuhan ekonomi lokal. Intensitas kunjungan wisata ke Danau Toba, juga memantik persepsi wisatawan terhadap kualitas produk dan layanan yang diperoleh selama berwisata. Disisi lain, masyarakat lokal yang terlibat sebagai pelaku usaha perjalanan wisata maupun sektor pendukung pariwisata juga memiliki kesan dan perspektif tersendiri terkait dengan program dan kebijakan pengembangan pariwisata Danau Toba. Irene et al. menunjukkan bahwa Danau Toba memiliki citra wisata sejarah dan alam sebagai atraksi yang memantik kunjungan wisata domestik dan mancanegara [31]. Disisi lain, Hidayat dan Nasution menunjukkan bahwa destinasi wisata Danau Toba juga dikenal sebagai Global Geopark Kaldera UNESCO [32]. Hal ini menunjukkan bahwa destinasi wisata Danau Toba memiliki atraksi yang perlu dioptimalkan dengan berbagai sarana dan prasarana yang berhubungan dengan aksesibilitas, akomodasi dan amenitas. Dengan demikian, wisatawan yang berkunjung ke destinasi wisata Danau Toba memiliki pengalaman yang menyenangkan, serta memiliki tingkat kepuasan yang tinggi terhadap produk dan layanan.

Penelitian ini menggunakan data teks ulasan pengunjung terhadap destinasi wisata Danau Toba, yang diperoleh dari website Tripadvisor. Terdapat 858 ulasan yang terpublikasi dengan klasifikasi rating sebagai berikut : 8 ulasan dengan rating sangat buruk; 22 ulasan dengan rating buruk; 81 ulasan dengan rating netral; 304 ulasan dengan rating baik; 443 ulasan dengan rating sangat baik. Setelah dilakukan proses scraping data teks dan diperoleh 424 ulasan wisatawan yang dapat dilanjutkan ke tahap pre-processing menggunakan teknik TF-IDF pada persamaan (6) dan persamaan (7). Hasil pre-processing data menunjukkan adanya 382 data ulasan yang dapat diklasifikasi menggunakan algoritma NBC dan DT sebagai model. Meskipun demikian, pada tahap data processing, perlu digunakan operator SMOTE Upsampling dimana proses perhitungannya dapat dilihat pada persamaan (8) dan persamaan (9) dengan pembagian data uji (70%) dan data latih (30%). Selanjutnya, pengujian model pertama menggunakan algoritma NBC menggunakan persamaan (1), persamaan (2), dan persamaan (3). Adapun, luaran hasil kalkulasi menggunakan algoritma NBC dapat dilihat dalam bentuk confusion matrix untuk memperoleh nilai akurasi, presisi, recall, dan f-measure yang diperoleh dari persamaan (10), persamaan (11), persamaan (12) dan persamaan (13). Perhitungan model NBC tidak dilakukan secara manual melainkan melalui aplikasi Rapidminer, sebagaimana gambar berikut.



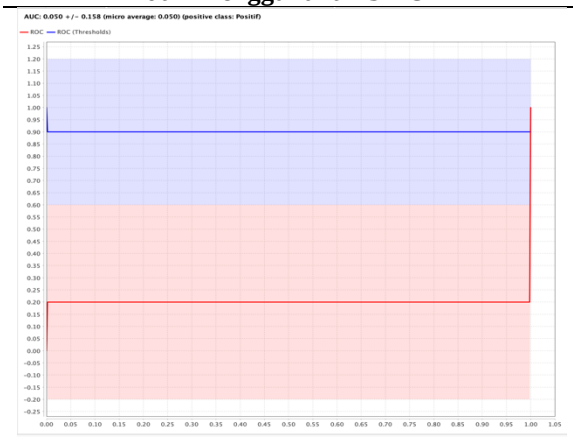
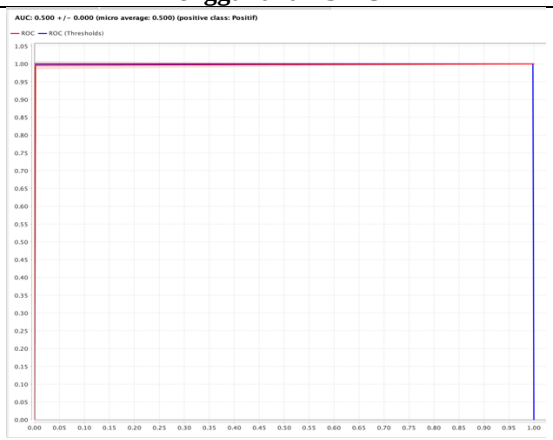
Gambar 5. Cross Folding Algoritma NBC dengan dan tanpa SMOTE

Gambar 5 menunjukkan visualisasi operator SMOTE UPsampling dan operator Cross Folding dalam pengujian model NBC. Hasil klasifikasi algoritma NBC tanpa menggunakan operator SMOTE UPsampling menunjukkan nilai akurasi (96,64%), presisi (97,01%), recall (99,61%), f-measure (98,29), Area Under Curve (0,050). Berbeda halnya dengna hasil klasifikasi algoritma NBC menggunakan operator SMOTE UPsampling yang menunjukkan nilai akurasi (99,81%), presisi (100%), recall (99,62%), f-measure (99,80%), Area Under Curve (0,500). Berdasarkan confusion matrix dapat diketahui bahwa hasil penggunaan operator SMOTE UPsampling menunjukkan nilai akurasi, presisi, recall, f-measure yang lebih baik. Meskipun demikian, nilai AUC 0,500 menunjukkan bahwa model tersebut tidak memiliki kemampuan klasifikasi yang acak. A'yuniyah et al. menunjukkan bahwa model klasifikasi yang dengan nilai AUC 0,9-1 dinilai sebagai model yang sangat baik serta memiliki kemampuan klasifikasi yang baik serta mampu membedakan kelas negatif dan positf [33]. Hal ini menunjukkan bahwa hasil pengolahan data ulasan menggunakan algoritma NBC, dengan jumlah data latih (30%) dan data uji (70%) menggunakan operator SMOTE UPsampling maupun tanpa menggunakan operator SMOTE UPsampling menunjukkan performa rendah sebagaimana tabel berikut.

Submitted: **26/06/2023**; Accepted: **31/07/2023**; Published: **31/07/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1551-1562
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6461

Tabel 1. AUC dan Confusion Matrix algoritma NBC (dengan dan tanpa menggunakan SMOTE)

Tidak Menggunakan SMOTE	Menggunakan SMOTE
	
PerformanceVector: accuracy: 96.64% +/- 1.18% (micro average: 96.63%) ConfusionMatrix: True: Negatif Positif Negatif: 0 1 Positif: 8 258 AUC (optimistic): 0.850 +/- 0.337 (micro average: 0.850) (positive class: Positif) AUC: 0.050 +/- 0.158 (micro average: 0.050) (positive class: Positif) AUC (pessimistic): 0.050 +/- 0.158 (micro average: 0.050) (positive class: Positif) precision: 97.01% +/- 1.58% (micro average: 96.99%) (positive class: Positif) ConfusionMatrix: True: Negatif Positif Negatif: 0 1 Positif: 8 258 recall: 99.63% +/- 1.17% (micro average: 99.61%) (positive class: Positif) ConfusionMatrix: True: Negatif Positif Negatif: 0 1 Positif: 8 258 f_measure: 98.29% +/- 0.60% (micro average: 98.29%) (positive class: Positif) ConfusionMatrix: True: Negatif Positif Negatif: 0 1 Positif: 8 258	PerformanceVector: accuracy: 99.81% +/- 0.61% (micro average: 99.81%) ConfusionMatrix: True: Negatif Positif Negatif: 259 1 Positif: 0 258 AUC (optimistic): 1.000 +/- 0.000 (micro average: 1.000) (positive class: Positif) AUC: 0.500 +/- 0.000 (micro average: 0.500) (positive class: Positif) AUC (pessimistic): 0.996 +/- 0.012 (micro average: 0.996) (positive class: Positif) precision: 100.00% +/- 0.00% (micro average: 100.00%) (positive class: Positif) ConfusionMatrix: True: Negatif Positif Negatif: 259 1 Positif: 0 258 recall: 99.62% +/- 1.22% (micro average: 99.61%) (positive class: Positif) ConfusionMatrix: True: Negatif Positif Negatif: 259 1 Positif: 0 258 f_measure: 99.80% +/- 0.62% (micro average: 99.81%) (positive class: Positif) ConfusionMatrix: True: Negatif Positif Negatif: 259 1 Positif: 0 258

Tabel 1 menunjukkan bahwa nilai confusion matrix algoritma NBC memiliki nilai akurasi, presisis, recall, dan f-measure yang tinggi, meskipun demikian nilai AUC menunjukkan bahwa model yang digunakan dikategorikan sangat buruk atau tidak memiliki kemampuan untuk membedakan kelas negatif dan positif. Permatasari dan Irhamah menunjukkan bahwa algoritma dengan nilai AUC di atas 0,7 dapat digunakan sebagai model klasifikasi, sedangkan nilai AUC 0,5 dinilai memiliki performa yang tidak optimal [34]. Disisi lain, Nuraeni menunjukkan bahwa performa algoritma dapat diklasifikasikan sebagai excellent classification apabila memiliki nilai 0,9 sampai dengan 1 [35]. Hal ini menunjukkan bahwa nilai AUC NBC perlu dibandingkan dengan nilai AUC algoritma DT untuk mengevaluasi model dengan performa terbaik. Mempertimbangkan hal tersebut, perlu diperbandingkan hasil klasifikasi algoritma NBC dengan DT menggunakan SMOTE UPsampling, maupun tanpa menggunakan SMOTE UPsampling.

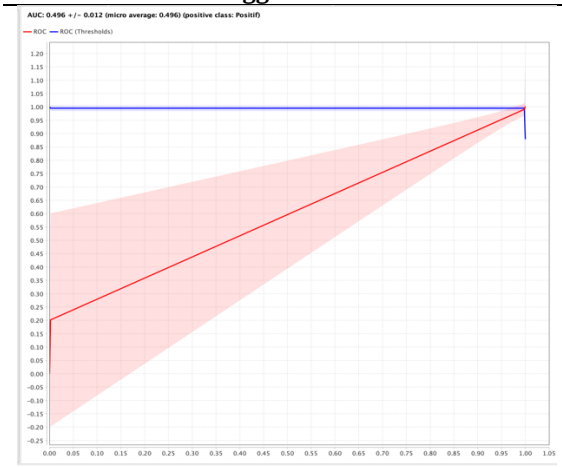
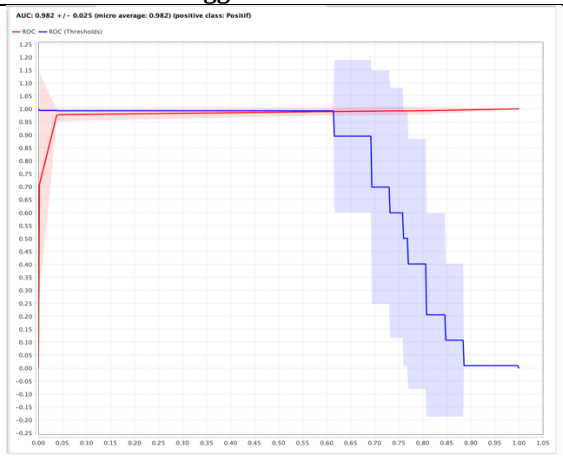
Berdasarkan hasil klasifikasi sentimen negatif dan positif menggunakan algoritma DT tanpa operator SMOTE UPsampling, dapat diketahui confusion matrix dengan nilai akurasi (96,27%), presisi (96,98%), recall

Submitted: **26/06/2023**; Accepted: **31/07/2023**; Published: **31/07/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1551-1562
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6461

(99,23%), f-measure (98,08%) dan nilai AUC (0,496). Berbeda halnya dengah hasil klasifikasi algoritma DT menggunakan operator SMOTE UPsampling yang menunjukkan nilai akurasi (98,27%), presisi (98,83%), recall (97,71%), f-measure (98,26%), dan nilai AUC (0,982). Hal ini menunjukkan bahwa algoritma DT menggunakan operator SMOTE UPsampling memiliki performa yang lebih baik dibandingkan tanpa menggunakan operator SMOTE UPsampling. Adapun, hasil klasifikasi algoritma DT menggunakan aplikasi Rapidminer sebagaimana persamaan (4) dan persamaan (5) dapat dilihat pada confusion matrix yang ditampilkan dalam table berikut.

Tabel 2. AUC dan Confusion Matrix algoritma DT (dengan dan tanpa menggunakan SMOTE)

Tidak Menggunakan SMOTE	Menggunakan SMOTE
	
PerformanceVector: accuracy: 96.27% +/- 3.02% (micro average: 96.25%) ConfusionMatrix: True: Negatif Positif Negatif: 0 2 Positif: 8 257 AUC (optimistic): 0.892 +/- 0.208 (micro average: 0.892) (positive class: Positif) AUC: 0.496 +/- 0.012 (micro average: 0.496) (positive class: Positif) AUC (pessimistic): 0.100 +/- 0.211 (micro average: 0.100) (positive class: Positif) precision: 96.98% +/- 1.60% (micro average: 96.98%) (positive class: Positif) ConfusionMatrix: True: Negatif Positif Negatif: 0 2 Positif: 8 257 recall: 99.23% +/- 2.43% (micro average: 99.23%) (positive class: Positif) ConfusionMatrix: True: Negatif Positif Negatif: 0 2 Positif: 8 257 f_measure: 98.08% +/- 1.60% (micro average: 98.09%) (positive class: Positif) ConfusionMatrix: True: Negatif Positif Negatif: 0 2 Positif: 8 257	PerformanceVector: accuracy: 98.27% +/- 2.12% (micro average: 98.26%) ConfusionMatrix: True: Negatif Positif Negatif: 256 6 Positif: 3 253 AUC (optimistic): 0.994 +/- 0.013 (micro average: 0.994) (positive class: Positif) AUC: 0.982 +/- 0.025 (micro average: 0.982) (positive class: Positif) AUC (pessimistic): 0.970 +/- 0.038 (micro average: 0.970) (positive class: Positif) precision: 98.83% +/- 1.88% (micro average: 98.83%) (positive class: Positif) ConfusionMatrix: True: Negatif Positif Negatif: 256 6 Positif: 3 253 recall: 97.71% +/- 2.68% (micro average: 97.68%) (positive class: Positif) ConfusionMatrix: True: Negatif Positif Negatif: 256 6 Positif: 3 253 f_measure: 98.26% +/- 2.14% (micro average: 98.25%) (positive class: Positif) ConfusionMatrix: True: Negatif Positif Negatif: 256 6 Positif: 3 253

Tabel 2 merupakan visualisasi AUC dan confusion matrix algoritma DT dengan dan tanpa menggunakan operator SMOTE UPsampling. Model terbaik dalam penerapan metode klasifikasi diperoleh dari penggunaan operator SMOTE UPsampling pada algoritma DT yang menunjukkan nilai AUC sebesar 0,982 (98,2%) sehingga dapat dikagetorikan sebagai excellent classification atau model dengan kemampuan klasifikasi yang baik. Hal ini menunjukkan bahwa pada tahap pemodelan, algoritma DT menggunakan SMOTE dapat menghasilkan performa

Submitted: **26/06/2023**; Accepted: **31/07/2023**; Published: **31/07/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1551-1562
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6461

yang lebih baik. Dengan demikian, dapat diketahui bahwa mayoritas data ulasan pengunjung mengandung unsur positif yang dominan pada aspek atraksi Danau Toba dan Pulau Samosir.

Dalam konteks kepariwisataan, perencanaan pengembangan destinasi wisata perlu melibatkan masyarakat lokal agar tidak menjadi faktor ketimpangan ekonomi-pariwisata yang berdampak pada kondisi sosial-budaya masyarakat dan lingkungan sekitar. Selain itu, komponen atraksi, aksesibilitas, akomodasi dan amentias perlu dikembangkan secara seimbang agar mendorong minat berkunjung, maupun motivasi berkunjung kembali ke Danau Toba. Penelitian ini menunjukkan bahwa mayoritas pengulas menulis pengalaman positif dan kesan yang baik melalui website Tripavisor, namun tidak berarti ulasan yang negatif dalam pengalaman perorangan diabaikan, tanpa ada tanggapan atau tindakan untuk membenahi kesalahan. Berdasarkan data teks ulasan dengan klasifikasi negatif dapat diketahui adanya penekanan pada kualitas layanan yang buruk serta minimnya sarana dan prasarana pendukung. Hal ini berarti bahwa manajemen operasional destinasi wisata memiliki kelemahan dalam upaya pengendalian atau pengawasan kinerja karyawan dalam memberikan pelayanan prima, oleh sebab itu diperlukan program pelatihan untuk meningkatkan kualitas Sumber Daya Manusia (SDM) di bidang hospitality sesuai standar pelayanan prima. Selanjutnya, pembangunan infrastruktur pariwisata perlu mempertimbangkan kebutuhan wisatawan, kondisi eksisting lingkungan dan pola penghidupan masyarakat lokal. Dengan demikian, pada tahap deployment, luaran penelitian dapat digunakan sebagai pertimbangan pengambilan kebijakan dan penetapan program priorias yang berhubungan dengan peningkatan kualitas SDM serta pembangunan infrastruktur pariwisata.

4. KESIMPULAN

Hasil penelitian ini menunjukkan bahwa proses klasifikasi algoritma DT menggunakan operator SMOTE UPsampling menunjukkan performa yang baik dengan nilai akurasi (98,27%), presisi (98,83%), recall (97,71%), f-measure (98,26%), dan nilai AUC (0,982). Apabila dibandingkan dengan algoritma NBC, model yang paling relevan dengan konteks dataset ulasan pengunjung Danau Toba ialah model algoritma DT menggunakan operator UPsampling. Disisi lain, terdapat lima kata populer dalam data ulasan pengunjung Danau Toba yang menunjukkan sorotan wisatawan pada atraksi, aksesiblitas, akomodasi dan amentias pendukung pariwisata. Berdasarkan lima kata populer terdapat kata-kata sebagai berikut : danau (329), toba (254), pulau (156), samosir (154), tempat (143). Hal ini menunjukkan bahwa atraksi menjadi aspek penting dalam persepsi wisatawan, sehingga perlu dikelola dengan optimal. Dengan demikian, pengelola perlu menyeimbangkan program pengembangan atraksi wisata dengan sarana dan prasarana pendukung aktivitas wisata, agar mendukung kepuasan berkunjung dan memantik motivasi berkunjung kembali.

UCAPAN TERIMA KASIH

Terima kasih disampaikan kepada Lembaga Penelitian dan Pengabdian kepada Masyarakat, Program Studi Pariwisata, Program Studi Sistem Informasi, Fakultas Ilmu Administrasi Bisnis dan Ilmu Komunikasi, Fakultas Teknik, Universitas Katolik Indonesia Atma Jaya, atas dukungan dalam penelitian hingga publikasi hasil penelitian ini.

REFERENCES

[1] F. T. Meturan, M. Idris Taking, and R. Latief, “Analisis Ketersediaan Prasaran Dan Fasilitas Penunjang Pengembangan Objek Wisata Pantai Liang Kecamatan Salahutu Kabupaten Maluku Tengah,” J. Urban Plan. Stud., vol. 2, no. 1, pp. 85–95, 2021, doi: 10.35965/jups.v2i1.33.

[2] S. A. Azzahra and A. Wibowo, “Analisis Sentimen Multi-Aspek Berbasis Konversi Ikon Emosi dengan Algoritme Naïve Bayes untuk Ulasan Wisata Kuliner Pada Web Tripadvisor,” J. Teknol. Inf. dan Ilmu Komput., vol. 7, no. 4, pp. 737–744, 2020, doi: 10.25126/jtiik.2020731907.

[3] D. Riadi, L. A. Permadi, and W. Retnowati, “Pengaruh Kualitas Pelayanan Terhadap Minat Berkunjung Kembali ke Desa Wisata Hijau Bilebante yang Dimediasi Oleh Kepuasan Wisatawan,” J. Ris. Pemasar., vol. 2, no. 2, pp. 38–49, 2023.

[4] Y. Christian and K. O. Y. R. Qi, “Penerapan K-Means pada Segmentasi Pasar untuk Riset Pemasaran pada Startup Early Stage dengan Menggunakan CRISP-DM,” JURIKOM (Jurnal Ris. Komputer), vol. 9, no. 4, pp. 966–973, 2022, doi: 10.30865/jurikom.v9i4.4486.

[5] D. A. Munawwaroh and A. H. Primandari, “Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Ibu Hamil Berpotensi Gizi Kurang,” Delta J. Ilm. Pendidik. Mat., vol. 10, no. 2, pp. 367–380, 2022, doi: 10.30871/jaic.v5i2.3200.

[6] Y. Harwani, “Constructing Brand Personality from the TripAdvisor Online Reviews,” Int. J. Bus. Manag. Technol., vol. 5, no. 4, pp. 152–157, 2021, doi: https://dx.doi.org/10.5281/zenodo.7672647.

[7] P. Rita, R. Ramos, M. T. Borges-Tiago, and D. Rodrigues, “Impact of the rating system on sentiment and tone of voice: A Booking.com and TripAdvisor comparison study,” Int. J. Hosp. Manag., vol. 104, no. 2, pp. 1–12, 2022, doi: 10.1016/j.ijhm.2022.103245.

[8] P. P. Dewi, I. P. Utama, and I. A. P. Widawati, “Peran Brand Image Situs Tripadvisor Memediasi Pengaruh eWoM tehradap Niat Beli Kamar di Kabupaten Badung,” TULIP Tulisan Ilm. Pariwisata, vol. 5, no. 2, pp. 75–81, 2022.

[9] A. Lynn, M. T. J. T, A. Lianina, and F. A. Madrilejos, “A Narrative Analysis on Tripadvisor Reviews of Guest

Submitted: **26/06/2023**; Accepted: **31/07/2023**; Published: **31/07/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1551-1562
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6461

Satisfaction in Conrad Manila as a Quarantine Facility 2020-2021," Int. J. Manag. Commer. Innov., vol. 10, no. 1, pp. 435–446, 2022, doi: https://doi.org/10.5281/zenodo.7027432.
[10] A. Minkwitz, "Tripadvisor as a source of data in the planning process of tourism development on a local scale," Turyzm/Tourism, vol. 28, no. 2, pp. 49–55, 2018, doi: 10.2478/tour-2018-0014.
[11] D. A. Pramudita and Bagus Sumargo, "Pengelompokan Pengguna Internet dengan Metode K-Means Clustering," J. Stat. dan Apl., vol. 3, no. 1, pp. 1–12, 2019, doi: 10.21009/jsa.03101.
[12] R. Santoso, H. A. Munawi, and D. Sukmawati, "Perkembangan Teknologi Informasi dan Telekomunikasi Terhadap Perubahan Perilaku Masyarakat," in Conference on Research and Community Services, 2019, pp. 586–592.
[13] Barrie Goldsmith, "Negative Feedback on Tripadvisor: A Hotel's Nightmare," J. Tour. Hosp. Manag., vol. 4, no. 3, pp. 135–138, 2016, doi: 10.17265/2328-2169/2016.06.004.
[14] W. Khofifah, D. N. Rahayu, and A. M. Yusuf, "Analisis Sentimen Menggunakan Naive Bayes Untuk Melihat Review Masyarakat Terhadap Tempat Wisata Pantai Di Kabupaten Karawang Pada Ulasan Google Maps," J. Interkom J. Publ. Ilm. Bid. Teknol. Inf. dan Komun., vol. 16, no. 4, pp. 28–38, 2022, doi: 10.35969/interkom.v16i4.192.
[15] G. K. Pati and E. Umar, "Analisis Sentimen Komentar Pengunjung Terhadap Tempat Wisata Danau Weekuri Menggunakan Metode Naive Bayes Classifier Dan K-Nearest Neighbor," J. Media Inform. Budidarma, vol. 6, no. 4, pp. 2309–2315, 2022, doi: 10.30865/mib.v6i4.4635.
[16] H. Christanto et al., "Analisis Perbandingan Decision Tree , Support Vector Machine , dan Xgboost dalam Mengklasifikasi Review Hotel Trip Advisor," J. Teknol. Inform. dan Komput. MH. Thamrin, vol. 9, no. 1, pp. 306–319, 2023.
[17] A. A. Arifiyanti, M. Fuad, P. Fikri, and B. Utomo, "Analisis Sentimen Ulasan Pengunjung Objek Wisata Gunung Bromo pada Situs Tripadvisor," Explor. J. Sist. Inf. dan Telemat., vol. 13, no. 1, pp. 32–37, 2022.
[18] O. Somantri and Dairoh, "Analisis Sentimen Penilaian Tempat Tujuan Wisata Kota Tegal Berbasis Text Mining," JEPIN J. Edukasi dan Penelit. Inform., vol. 5, no. 2, pp. 191–196, 2019.
[19] F. Nurhuda, S. W. Sihwi, and A. Doewes, "Analisis Sentimen Masyarakat Terhadap Pilpres 2019 Berdasarkan Opini Dari Twitter Menggunakan Metode Naive Bayes Classifier," J. ITSMART, vol. 2, no. 2, pp. 35–42, 2013, doi: 10.51519/journalcisa.v1i3.45.
[20] M. F. Asshiddiqi and K. M. Lhaksmana, "Perbandingan Metode Decision Tree dan Support Vector Machine untuk Analisis Sentimen pada Instagram Mengenai Kinerja PSSI," in e-Proceeding of Engineering, 2020, vol. 7, no. 3, pp. 9936–9948.
[21] S. W. U. Vitandy, A. A. Supianto, and F. A. Bachtiar, "Analisis Sentimen Evaluasi Kinerja Dosen menggunakan Term Frequency- Inverse Document Frequency dan Naïve Bayes Classifier," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 3, no. 6, pp. 6080–6088, 2019, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/5645
[22] R. Kosasih and A. Alberto, "Analisis Sentimen Produk Permainan Menggunakan Metode TF-IDF Dan Algoritma K-Nearest Neighbor," InfoTekJar J. Nas. Inform. dan Teknol. Jar., vol. 6, no. 1, pp. 134–139, 2021.
[23] M. Y. Ardiansyah, M. A. Fauzi, and S. Adinugroho, "Penerapan Term Frequency-Modified Inverse Document Frequency pada Analisis Sentimen Ulasan Barang menggunakan Metode Learning Vector Quantization," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 3, no. 6, pp. 5592–5598, 2019, [Online]. Available: http://j-ptiik.ub.ac.id
[24] M. I. Alfarizi, L. Syafaah, and M. Lestandy, "Emotional Text Classification Using TF-IDF (Term Frequency-Inverse Document Frequency) And LSTM (Long Short-Term Memory)," JUITA J. Inform., vol. 10, no. 2, p. 225, 2022, doi: 10.30595/juita.v10i2.13262.
[25] E. B. S. Rayhan Rahmanda, "JURNAL RESTI Word2Vec on Sentiment Analysis with Synthetic Minority Oversampling," J. RESTI, vol. 5, no. 2, pp. 599–605, 2022.
[26] R. A. Barro, I. D. Sulvianti, and M. Afendi, "Penerapan Synthetic Minority Oversampling Technique (Smote) Terhadap Data Tidak Seimbang Pada Pembuatan Model Komposisi Jamu," Xplore J. Stat., vol. 1, no. 1, pp. 1–6, 2013.
[27] Y. E. Kurniawati, "Class Imbalanced Learning Menggunakan Algoritma Synthetic Minority Over-sampling Technique – Nominal (SMOTE-N) pada Dataset Tuberculosis Anak," J. Buana Inform., vol. 10, no. 2, pp. 134–143, 2019, doi: 10.24002/jbi.v10i2.2441.
[28] N. L. W. S. R. Ginantra, C. P. Yanti, G. D. Prasetya, I. B. G. Sarasvandana, and I. K. A. G. Wiguna, "Analisis Sentimen Ulasan Villa di Ubud Menggunakan Metode Naive Bayes, Decision Tree, dan k-NN," Janapati, vol. 11, no. 3, pp. 205–216, 2022.
[29] J. Pardosi, R. Sibarani, N. C. Bangun, and I. M. Putra, "Peran Sumber Daya Manusia Transportasi Penyeberangan dalam Meningkatkan Pelayanan Pariwisata di Danau Toba," War. Penelit. Perhub., vol. 33, no. 2, pp. 113–122, 2021.
[30] N. A. Wulandari, D. S. Kartini, and N. Y. Yuningsih, "Akselerasi Pengembangan Destinasi Wisata Danau Toba (Studi Realisasi Prinsip Good Governance Pada Badan Pelaksana Otorita Danau Toba)," J. MODERAT, vol. 7, no. 3, pp. 512–533, 2021.
[31] O. Irena, D. Christie, and S. Thio, "Persepsi Masyarakat Terhadap Citra Destinasi Dari Candi Borobudur, Mandalika, Labuan Bajo, dan Danau Toba," J. Hosp. dan Manaj. Jasa, vol. 7, no. 2, pp. 1–23, 2019.
[32] T. Wal hidayat and I. Nasution, "Persepsi Publik Tentang Destinasi Pariwisata Danau Toba Sebagai Global Geopark Kaldera UNESCO," Publikauma J. Adm. Publik Univ. Medan Area, vol. 7, no. 2, pp. 88–102, 2019, doi: 10.31289/publika.v7i2.2943.
[33] Q. A'yuniyah et al., "Implementasi Algoritma Naïve Bayes Classifier (NBC) untuk Klasifikasi Penyakit Ginjal Kronik," J. Sist. Komput. dan Inform., vol. 4, no. 1, pp. 72–76, 2022, doi: 10.30865/json.v4i1.4781.
[34] R. I. Permatasari et al., "Analisis Sentimen Film pada Twitter Berbahasa Indonesia Menggunakan Ensemble Features dan Naïve Bayes," J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya, vol. 2, no. 11, pp. 5921–5927, 2018.
[35] N. Nuraeni, "Penentuan Kelayakan Kredit Dengan Algoritma Naïve Bayes Classifier: Studi Kasus Bank Mayapada Mitra Usaha Cabang PGC," J. Tek. Komput., vol. 3, no. 1, pp. 9–15, 2017, [Online]. Available: https://ejournal.bsi.ac.id/ejurnal/index.php/jtk/article/view/1337