Building of Informatics, Technology and Science (BITS)
Volume 4, No 3, Desember 2022 Page: 1634−1646
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i3.2766

Analisis Performa Algoritma NBC, DT, SVM dalam Klasifikasi Data Ulasan Pengunjung Candi Borobudur Berbasis CRISP-DM

Yerik Afrianto Singgalen*

Fakultas Ilmu Administrasi Bisnis dan Komunikasi, Pariwisata, Universitas Katolik Indonesia Atma Jaya, Jakarta, Indonesia
Email: yerik.afrianto@atmajaya.ac.id
Email Penulis Korespondensi: yerik.afrianto@atmajaya.ac.id
Submitted: **26/12/2022**; Accepted: **30/12/2022**; Published**: 30/12/2022**

Abstrak−Pendekatan Analisis sentimen pengunjung ke destinasi wisata Candi Borobudur di Indonesia dapat diklasifikasi menggunakan berbagai algoritma untuk mendapatkan hasil yang optimal. Performa algoritma yang baik dapat dilihat dari nilai confusion matrix (accuracy, precision, recall), nilai Area Under Curve (AUC) maupun Receiver Operating Charachteristic (ROC). Penelitian ini menggunakan algoritma Naïve Bayes Classifier (NBC), Decision Tree (DT), dan Support Vector Machine (SVM) terhadap 3850 data teks yang diperoleh dari website Tripadvisor, khususnya ulasan pengunjung Candi Borobudur. Metode yang digunakan mengacu pada Cross-Industry Standard Process for Data Mining (CRISP-DM) untuk optimalisasi produk dan layanan destinasi wisata dengan memerhatikan enam tahapan sebagai berikut business understanding, data understanding, data preparation, modeling, evaluation dan deployment. Hasil penelitian ini menunjukkan bahwa hasil analisis performa algotirma NBC menunjukkan adanya perubahan nilai confusion matrix pada nilai akurasi dari 98.73% menjadi 95.6%, nilai presisi berubah dari 98.72% menjadi 98.97%, nilai recall juga berubah dari 100% menjadi 96.54%. Adapun, nilai Area Under Curve (AUC) juga mengalami perubahan dari 0,500 (50%) menjadi 0,693 (69,35%). Selain itu, hasil evaluasi performa algoritma DT menunjukkan perubahan nilai confusion matrix pada nilai akurasi dari 97.55% menjadi 94.40%, nilai presisi meningkat dari 97.63% menjadi 91.86%, nilai recall juga berubah dari 99.90% menjadi 99.47%. Adapun, nilai Area Under Curve (AUC) juga mengalami perubahan dari 0,591 (59,1%) menjadi 0,932 (93,2%). Adapun, hasil evaluasi performa algoritma SVM menunjukkan perubahan nilai confusion matrix pada nilai akurasi dari 98.73% menjadi 99.41%, nilai presisi berubah dari 98.72% menjadi 100%, nilai recall juga berubah dari 100% menjadi 99.01%. Adapun, nilai Area Under Curve (AUC) juga mengalami perubahan dari 0,961 (96,1%) menjadi 1,00 (100%). Selain itu, hasil uji T menunjukkan bahwa algoritma SVM lebih dominan dibandingkan dengan algoritma lainnya, dimana nilai uji T algoritma SVM sebesar 0,994 jika dibandingkan dengan nilai uji T algoritma DT sebesar 0,944 dan nilai uji T algoritma NBC sebesar 0,98. Berdasarkan nilai Receiver Operating Characteristic (ROC) dapat diketahui bahwa algoritma DT juga menunjukkan performa yang baik, selain SVM. Hal ini menunjukkan bahwa dalam proses analisis sentimen pengunjung ke Candi Borobudur, algoritma yang dapat digunakan ialah SVM.

Kata Kunci: Analisis Sentimen,;Naïve Bayes Classifier; Decision Tree; Support Vector Machine

Abstract−The approach of visitor sentiment analysis to Borobudur Temple tourist destinations in Indonesia can be classified using various algorithms to get optimal results. Good algorithm performance can be seen from the confusion matrix (accuracy, precision, recall) value, Area Under Curve (AUC) value, and Receiver Operating Characteristic (ROC). This study used the Naïve Bayes Classifier (NBC), Decision Tree (DT), and Support Vector Machine (SVM) algorithms against 3850 text data obtained from the Tripadvisor website, especially reviews of Borobudur Temple visitors. The method refers to the Cross-Industry Standard Process for Data Mining (CRISP-DM) for optimizing tourist destination products and services by paying attention to six stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The results of this study show that the results of NBC's algorithm performance evaluation can be seen to have a change in the confusion matrix value at the accuracy value from 98.73% to 95.6%, the precision value changed from 98.72% to 98.97%, the recall value also changed from 100% to 96.54%. In addition, the Area Under Curve (AUC) of NBC also changed from 0.500 (50%) to 0.693 (69.35%). In addition, the results of the DT algorithm performance evaluation showed a change in the confusion matrix value at the accuracy value from 97.55% to 94.40%, the precision value increased from 97.63% to 91.86%, the recall value also changed from 99.90% to 99.47%. The Area Under Curve (AUC) of DT value also changed from 0.591 (59.1%) to 0.932 (93.2%). The results of the SVM algorithm performance evaluation showed a change in the confusion matrix value at the accuracy value from 98.73% to 99.41%; the precision value changed from 98.72% to 100%, and the recall value also changed from 100% to 99.01%. The Area Under Curve (AUC) of the SVM value also changed from 0.961 (96.1%) to 1.00 (100%). In addition, the T-test results show that the SVM algorithm is more dominant compared to other algorithms, where the SVM algorithm T-test value is 0.994 compared to the DT algorithm T-test value of 0.944 and the NBC algorithm T-test value of 0.98. Based on the Receiver Operating Characteristic (ROC) value, it can be seen that the DT algorithm also shows good performance in addition to SVM. It indicates that in analyzing the sentiment of visitors to Borobudur Temple, the best-recommended algorithm is the Support Vector Machine.

Keywords: Sentiment Analysis; Naïve Bayes Classifier; Decision Tree; Support Vector Machine

1. PENDAHULUAN

Perkembangan teknologi digital telah mendorong transformasi digital di berbagai negara untuk mengoptimalkan teknologi informasi sebagai penggerak pertumbuhan ekonomi dan sosial. Sektor pariwisata merupakan salah satu sektor yang mengalami perkembangan akibat transformasi digital di Indonesia [1]. Perkembangan pemasaran digital di sektor pariwisata memantik pertumbuhan aplikasi serta pengguna sistem informasi pariwisata untuk mengakses layanan informasi digital yang berhubungan dengan pariwisata [2]. Selain itu, Suasih et al., menunjukan bahwa salah satu faktor pendorong transformasi digital di sektor pariwisata ialah partisipasi pelaku Usaha Mikro, Kecil dan Menengah (UMKM) dalam menggunakan teknologi digital sebagai media pemasaran produk dan layanan di bisnis pariwisata [3]. Hal ini menunjukkan bahwa transformasi digital telah mendorong perkembangan teknologi informasi

Building of Informatics, Technology and Science (BITS)
Volume 4, No 3, Desember 2022 Page: 1634−1646
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i3.2766

sehingga masyarakat dan pemangku kepentingan di sektor pariwisata, perlu beradaptasi dengan perubahan *trend* pemasaran agar dapat menyesuaikan preferensi wisatawan [4]. Dengan demikian dapat diektahui bahwa data digital menjadi aset yang berperan penting dalam mendukung efektifitas dan efisiensi pemasaran produk dan layanan di bisnis pariwisata.

Beberapa studi tentang transformasi digital dan penambangan data (*data mining*) menunjukkan bahwa perusahaan yang mengadopsi teknologi informasi dalam proses bisnis, dapat meningkatkan kinerja karyawan serta mempercepat pencapaian target perusahaan secara efektif dan efisien [4]. Muttaqin et al., menunjukkan bahwa salah satu pendekatan pengolahan data teks yang dilakukan melalui proses penambangan data digital ialah *Cross Industry Standard Process for Data Mining* (CRISP-DM) [5]. Disisi lain, Felix et al., menunjukkan bahwa pendekatan *data mining* menekankan pada pengolahan data digital untuk mendukung proses pengambilan keputusan dalam meningkatkan manajemen perusahaan atau organisasi [6]. Dalam konteks pariwisata, penambangan data teks melalui pendekatan analisis sentimen dapat digunakan untuk memperoleh informasi terkait segmentasi pasar pariwisata maupun preferensi wisatawan terhadap produk dan layanan [7]. Hal ini menunjukkan bahwa data digital berupa teks, gambar, video, dan audio dapat dikelola untuk mengoptimalkan pemasaran destinasi pariwisata maupun bisnis pariwisata lainnya. Dengan demikian, perlu dilakukan analisis lebih lanjut tentang pendekatan analisis sentimen di bidang pariwisata.

Analisis sentimen dibutuhkan untuk memperoleh gambaran tentang kepuasan konsumen terkait dengan produk yang dikonsumsi atau layanan yang diterima ketika proses pembelian. Dalam konteks pariwisata, analisis sentimen wisatawan dibutuhkan untuk mengidentifikasi faktor-faktor yang memengaruhi kepuasan wisatawan ketika berkunjung ke suatu destinasi wisata. Metode CIRSP merupakan salah satu pendekatan yang relevan digunakan dalam analisis sentimen untuk mengetahui preferensi wisatawan terkait dengan atraksi, akomodasi dan amenitas, aksesibilitas maupun *ancillary*. Dengan demikian, pengelola destinasi wisata maupun para pemangku kepentingan di sektor pariwisata mendapatkan informasi tentang preferensi wisatawan terhadap produk dan layanan pariwisata melalui platform *digital* serta dapat menetapkan strategi pengembangan kaulitas produk dan layanan, maupun strategi manajemen destinasi wisata yang sesuai dengan tema pengembangan dan permintaan wisatawan.

Proses analisis sentimen wisatawan ditentukan oleh jumlah data teks, sumber data dan performa algoritma. Christanto & Singgalen menunjukkan bahwa wisatawan memiliki preferensi terhadap produk dan layanan yang disediakan oleh restaurant di suatu destinasi wisata, sehingga perlu diidentifikasi sentimen positif dan negatif wisatawan agar pihak pengelola bisnis dapat mengambil keputusan yang tepat dalam meningkatkan kualitas produk maupun layanan terhadap wisatawan [8]. Singgalen berpendapat bahwa ulasan negatif dan positif wisatawan terhadap produk dan layanan pariwisata merupakan data digital yang berperan esensial dalam menentukan inovasi bisnis pariwisata di masa mendatang [9]. Disisi lain, Wicaksono et al., berpendapat bahwa jumlah, sumber data teks dan algoritma yang digunakan sangat memengaruhi hasil klasifikasi sentimen negatif maupun positif [10]. Hal ini menunjukkan bahwa aspek esensial dalam proses analisis sentimen di bidang pariwisata ialah sumber data (platform digital), jumlah data, dan algortima yang digunakan. Dengan demikian dapat diketahui bahwa luaran proses analisis sentimen bersifat kontekstual sesuai dengan sumber data, jumlah data dan algoritma yang digunakan dalam metode klasifikasi.

Beberapa studi terdahulu menunjukkan bahwa metode klasifikasi untuk analisis sentimen yang dapat digunakan ialah algoritma *Naïve Bayes Classifier* (NBC), *Decision Tree* (DT), *Support Vector Machine* (SVM). Wahyu et al., menunjukkan bahwa data teks dari media sosial dapat digunakan sumber data untuk melakukan analisis sentimen, meskipun demikian data teks yang dikumpulkan harus relevan dengan konteks pembahasan dan topik yang spesifik [11]. Selain itu, Rahma et al., menunjukkan bahwa algoritma *Naïve Bayes Classifier* (NBC) dapat digunakan dalam analisis sentimen terkait dengan destinasi wisata [12]. Disisi lain, Arifiyanti et al., menggunakan algoritma *Decision Tree* (DT) dalam mengklasifikasi data teks yang berasal dari kolom ulasan website Tripadvisor [13]. Selain itu, Pratama et al., menggunakan algoritma Support Vector Machine (SVM) dalam klasifikasi data teks untuk menganalisis sentimen wisatawan di suatu destinasi wisata [14]. Hal ini menunjukkan bahwa website Tripadvisor dapat dijadikan sebagai sumber data teks untuk menganalisis sentimen wisatawan menggunakan algoritma *Naïve Bayes Classifier* (NBC), *Decision Tree* (DT), dan *Support Vector Machine* (SVM). Dengan demikian, dapat diketahui performa algoritma yang lebih baik dalam mengklasifikasi sentimen wisatawan.

Studi tentang analisis sentimen menggunakan pendekatan *data mining* menunjukkan sejumlah aplikasi yang dapat menyediakan operator atau plugin untuk memudahkan proses pengolahan data, salah satunya ialah Rapidminer. Meskipun demikian, keseimbangan data juga diperlukan untuk mengingkatkan performa algoritma melalui nilai Area Under Curve (AUC). Barro et al., menunjukkan bahwa salah satu operator dalam aplikasi Rapidminer yang digunakan untuk mengatasi masalah ketidakseimbangan data operator *Synthetic Minority Oversampling Technique* (SMOTE) [15]. Disisi lain, Hartono et al., menunjukkan bahwa ketidakseimbangan data dapat diatasi dengan *Synthetic Minority Oversampling Technique* (SMOTE) [16]. Hal ini menunjukkan bahwa proses analisis sentimen dapat dilakukan menggunakan aplikasi Rapidminer. Selanjutnya, ketidakseimbangan data dalam proses pengolahan data dapat diatasi dengan menggunakan operator SMOTE Upsampling. Dengan demikian, alur analisis sentimen dapat dilakukan secara efektif dan efisien untuk memperoleh informasi yang berguna bagi peningkatan kualitas produk dan layanan di destinasi wisata.

Penelitian ini bertujuan melakukan analisis sentimen pengunjung Candi Borobudur yang diperoleh dari kolom ulasan website Tripadvisor menggunakan metode *Cross Industry Standard Process for Data Mining* (CRISP-DM)

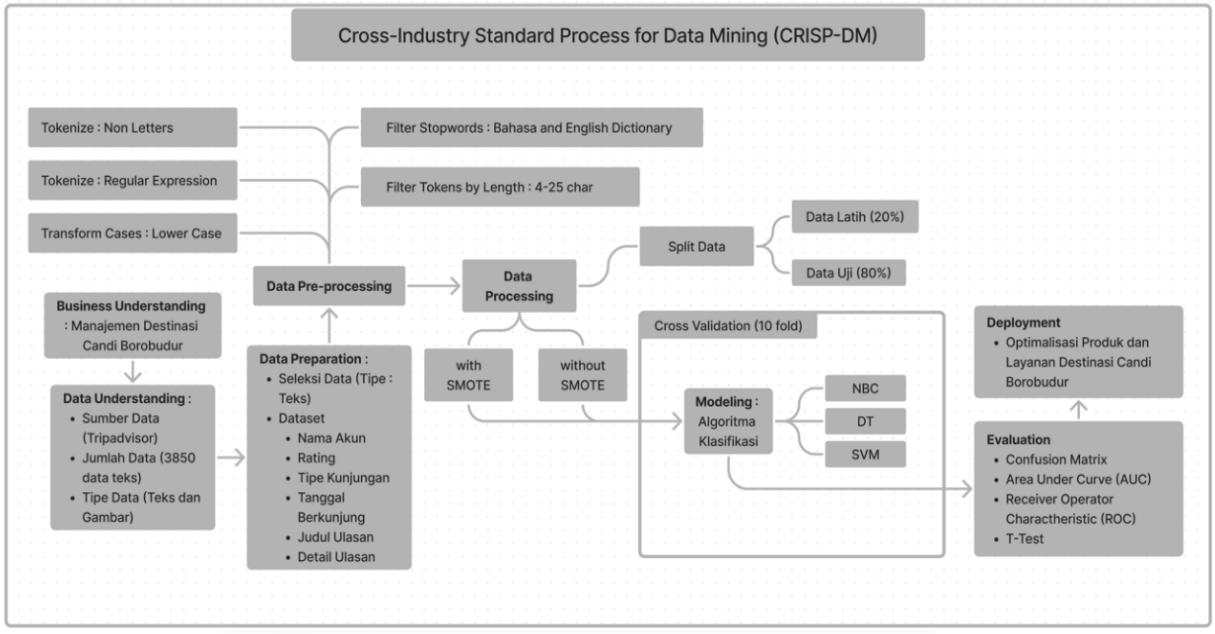
Building of Informatics, Technology and Science (BITS)
Volume 4, No 3, Desember 2022 Page: 1634−1646
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i3.2766

yang terbagi menjadi enam tahapan sebagai berikut *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation* dan *deployment*. Luaran penelitian ini dapat berupa rekomendasi pengembangan produk dan layanan destinasi Candi Borobudur berdasarkan jumlah kata yang telah diklasifikasi (*negative class* dan *positive class*). Selain itu, operator *Synthetic Minority Oversampling Technique* (SMOTE) digunakan untuk mengatasi masalah ketidakseimbangan data. Adapun, Algoritma yang digunakan dalam metode klasifikasi ialah algoritma *Naïve Bayes Classifier* (NBC), *Decision Tree* (DT), *Support Vector Machine* (SVM). Penelitian ini akan mengevaluasi performa masing-masing algoritma berdasarkan nilai *confusion matrix*, *Area Under Curve* (AUC), dan *Receiver Operating Characteristic* (ROC), and T-Test. Dengan demikian, algoritma dengan performa terbaik dapat dijadikan sebagai model dalam alur analisis sentimen data teks berikutnya.

2. METODOLOGI PENELITIAN

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

Metode yang digunakan dalam analisis sentimen pengunjung Candi Borobudur di website Tripadvisor ialah *Cross Industry Standard Process for Data Mining* (CRISP-DM) yang terbagi menjadi enam tahapan sebagai berikut *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation* dan *deployment*. Pada tahap *business understanding*, dilakukan studi literatur tentang sistem manajemen destinasi wisata, manajemen pemasaran destinasi berbasis website, segmentasi dan preferensi wisatawan terhadap produk dan layanan di destinasi wisata; tahap *data understanding*, dilakukan pengujian sistem validasi data ulasan oleh member website Tripadvisor, *screening* data ulasan berdasarkan nama akun, *rating*, tipe kunjungan, tanggal berkunjung, judul ulasan, dan detail ulasan, kemudian melakukan proses *scraping* data teks menggunakan aplikasi *webharvy*; tahap *data preparation*, dilakukan proses seleksi data yang akan dibersihkan (*tokenize, transform cases, filter stopwords, filter tokens by length*); tahap *data processing*, dilakukan pembagian data latih (20%) dan data uji (80%) dari 3850 data yang telah dihimpun, kemudian dilakukan pengujian algoritma menggunakan operator SMOTE Upsampling dan tanpa menggunakan SMOTE Upsampling untuk mengidentifikasi perbandingan nilai *Area Under Curve* (AUC); tahap *modeling*, dilakukan klasifikasi dataset menggunakan operator *cross validation* pada algoritma *Naïve Bayes Classifier* (NBC), *Decision Tree* (DT), *Support Vector Machine* (SVM); tahap *evaluation*, dilakukan analisis nilai *confusion matrix*, *Area Under Curve* (AUC), dan *Receiver Operating Characteristic* (ROC), and T-Test; tahap deployment, diberikan rekomendasi model dan luaran hasil analisis sentimen untuk optimalisasi produk dan layanan destinasi Candi Borobudur. Adapun, alur dari keseluruhan tahapan dapat dilihat pada gambar 1 berikut ini.



Gambar 1. Tahapan Penelitian sesuai CRISP-DM *Framework*

Gambar 1 merupakan keseluruhan tahapan penelitian berdasarkan kerangka kerja CRISP-DM dalam menganalisis analisis sentimen pengunjung Candi Borobudur di website Tripadvisor menggunakan algoritma *Naïve Bayes Classifier* (NBC), *Decision Tree* (DT), *Support Vector Machine* (SVM). Masing-masing algoritma memiliki keunggulan dan kelemahan sehingga perlu disesuaikan dengan karakteristik data yang akan dikelola untuk menghasilkan model yang sesuai. Hasil evaluasi algoritma dengan performa terbaik menunjukkan bahwa data tersebut dapat digunakan sebagai dasar untuk menghasilkan rekomendasi bagi pengelola destinasi wisata dalam mengoptimalkan produk dan layanan di Candi Borobudur, Indonesia.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 3, Desember 2022 Page: 1634−1646
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i3.2766

2.2 Algoritma *Naïve Bayes Classifier*

Naïve Bayes Classifier (NBC) memiliki keunggulan tersendiri dimana data diklasifikasi dengan probabilitas sederhana yang mengaplikasikan teorema bayes dengan asumsi ketidaktergantungan (independen) yang tinggi [17]. Penelitian ini didasarkan pada jumlah dataset yang digunakan untuk metode yang mempunyai performansi cepat dan akurat dalam pengklasifikasian. *Naïve Bayes Classifier* hanya membutuhkan data latih (*training data*) yang relatif kecil untuk menentukan estimasi parameter yang dibutuhkan dalam proses klasifikasi. Pada tahap klasifikasi, nilai kategori ditentukan dari data berdasarkan term yang muncul 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 jasa layanan pariwisata di Destinasi Candi Borobudur melalui website Tripadvisor.

2.3 Algoritma *Decision Tree*

Decision Tree (DT) merupakan salah satu metode klasifikasi populer yang sering digunakan secara praktis, model prediksi yang menggunakan struktur pohon untuk mencari dan membuat keputusan, serta memecahkan masalah dengan mempertimbangkan berbagai faktor di dalam lingkup masalah tersebut [18]. *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 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 adlah subset dari nilai Entropy. Puspita & Widodo berpendapat bahwa algoritma *Decision Tree* memiliki keunggulan dan kekurangan, dimana keunggulannya ialah konsep yang jelas dan mudah dipahami serta diimplementasikan menggunakan algoritma rekursif, sedangkan kekurangannya ialah tidak dapat diaplikasikan pada himpunan data yang sangat besar dan mudah mengalami overfit karena proses pelatihan *greedy* [19]. Dengan demikian perlu dilakukan pengujian algoritma SVM untuk membandingkan performa algoritma yang lebih baik.

2.4 Algoritma *Support Vector Machine*

Support Vector Machine (SVM) merupakan salah satu algoritma yang dapat digunakan untuk klasifikasi data menggunakan hyperplane [20]. Karim menjelaskan bahwa konsep SVM menitikberatkan pada *risk minimization,* yaitu estimasi fungsi dengan cara meminimalisir batas dari *generalization error,* sehingga SVM mampu mengatasi *overfitting* [21]. Adapun, fungsi regresi dari metode SVM adalah sebagai berikut.

$$f(x) = w^T \varphi(x) + b \quad (6)$$

Building of Informatics, Technology and Science (BITS)
Volume 4, No 3, Desember 2022 Page: 1634−1646
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i3.2766

Dimana w merupakan vector pembobot, $\varphi(x)$ merupakan sebuah fungsi yang memetakan x ke dalam suatu dimensi, dan b merupakan faktor bias. Selanjutnya, Nida et al., menunjukkan bahwa SVM mememiliki kelebihan dalam generalisasi data yang tinggi serta mampu menghasilkan model klasifikasi yang baik meskipun dilatih dengan data yang relatif sedikit. Meskipun demikian, sangat sulit diaplikasikan untuk himpunan data dengan sampel dan dimensi yang besar [22]. Hal ini menunjukkan bahwa SVM mampu menghasilkan performa yang baik meskipun dengan jumlah data yang relatif sedikit. Dengan demikian, perlu dilakukan pengujian algoritma SVM dengan dataset ulasan pengunjung Candi Borobudur.

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan, Persiapan dan Pengolahan Data di Rapidminer

Tripadvisor adalah platform digital berbasis web yang dapat digunakan sebagai forum informasi bagi wisatawan untuk memberikan rekomendasi wisata, serta akomodasi dan transportasi yang dapat digunakan wisatawan selama perjalanannya. Selain itu, Tripadvisor memiliki kolom ulasan tempat wisatawan dapat berbagi pengalaman dan menjelaskan kesan dan pesan setelah kunjungan mereka. Selain itu, emotikon dan fungsi penilaian tersedia dan dapat digunakan oleh pengulas untuk mengungkapkan pengalaman individu atau kelompok setelah mengunjungi suatu tujuan. Berdasarkan data Tripadvisor, terdapat 7.603 ulasan dengan peringkat yang berbeda-beda, yaitu sangat baik, sangat baik, rata-rata, buruk, dan buruk. Data teks yang diterima sekaligus adalah sebagai berikut.

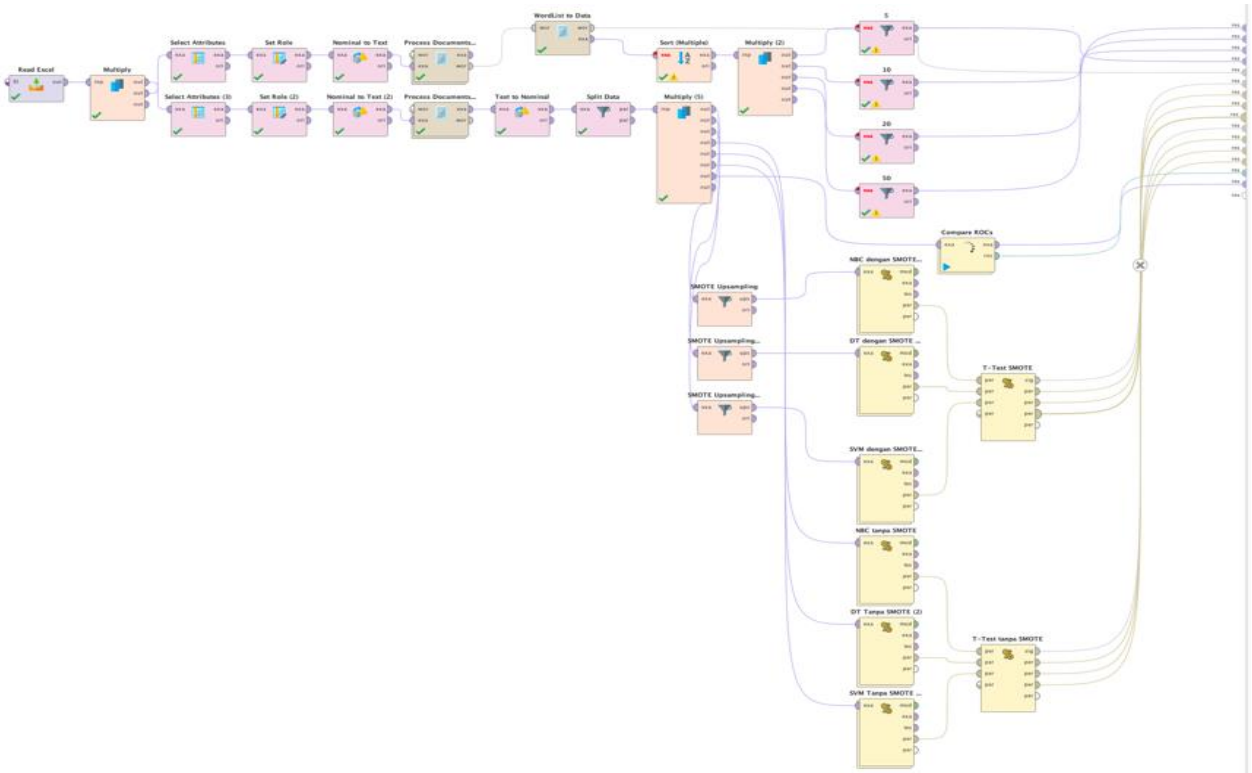
Sumber:*https://www.tripadvisor.com/Attraction_Review-g790291-d320054-Reviews-or10Borobudur_TempleBorobudur_Magelang_Central_Java_Java.html*.

Gambar 2. Data Ulasan Wisatawan di Halaman Website Tripadvisor (Candi Borobudur)

Gambar 2 memperlihatkan user interface halaman Candi Borobudur di website Tripadvisor dan hasil pengumpulan data, dikategorikan berdasarkan nama akun, ulasan, rating, bulan dan tahun pengajuan, jenis kunjungan, judul ulasan dan detail ulasan. Adapun proses verifikasi data berdasarkan identitas akun verifikasi (user account identity), rating (1-5), tanggal verifikasi (bulan:Januari-Desember) dan jenis kunjungan (bisnis, dengan pasangan, dengan keluarga, teman, sendirian). Berdasarkan data yang ditampilkan di situs web Tripadvisor, dikumpulkan 7.603 data ulasan dan 4.217 data ulasan berbahasa Inggris. Hal ini menunjukkan bahwa situs web Tripadvisor memuat beragam informasi ulasan, mulai dari bahasa hingga komponen produk dan layanan yang diulas. Namun, ulasan pengguna di situs Tripadvisor masih tidak terstruktur sehingga memerlukan pemrosesan data untuk memberikan informasi yang valid dan kredibel. Berdasarkan hal tersebut, penelitian ini berfokus pada data ulasan berbahasa Inggris, dengan mempertimbangkan peringkat produk dan layanan dalam ulasan wisatawan asing berbahasa Inggris. Pengumpulan data melalui website Tripadvisor dilakukan melalui aplikasi Webharvy. Konfigurasi untuk mempermudah proses pengambilan data dilakukan dalam beberapa langkah sebagai berikut: tambahkan tautan (*insert link*); pengaturan (configuration); membuat dan memperbarui informasi (*create and update data*); Atur tabel data (*setting data table*) pengaturan tempat/lokasi penyimpanan file (*setting path directory*); Pengumpulan dan ekspor data (*scraping and export data*).

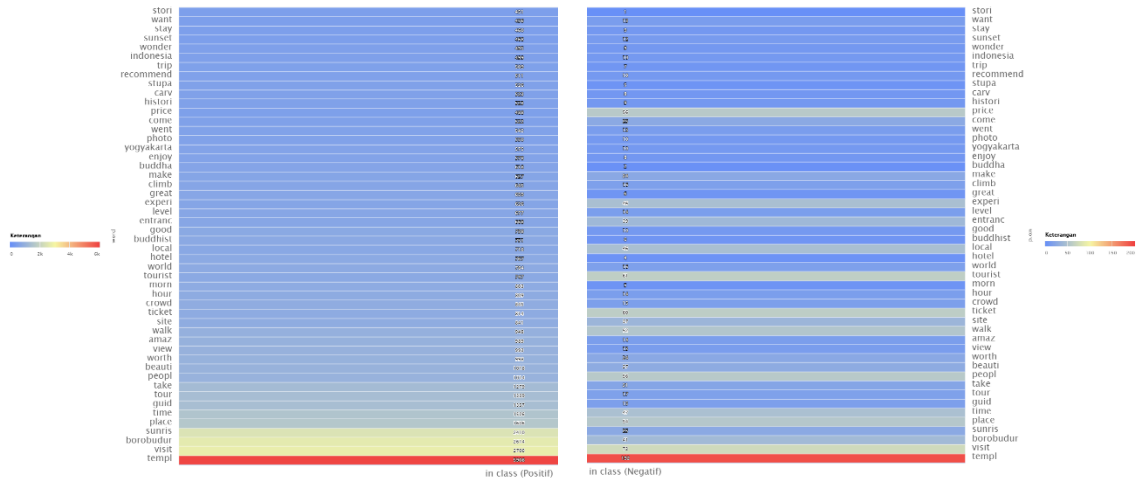
Langkah pra-pemrosesan diperlukan untuk menyiapkan data sehingga dapat dihitung dengan algoritma tertentu (NBC, DT, SVM). *Data preprocessing* sangat berguna untuk menyaring dan mengatur data agar tidak identik (duplikat) atau kosong (*missing data*). Berdasarkan hasil scraping di Tripadvisor, pada langkah pra-pemrosesan harus disiapkan 4.217 data ulasan sebelum proses perhitungan dilakukan oleh algoritma klasifikasi NBC, DT dan SVM, dimana 852 data ulasan tidak menyertakan ulasan terperinci, 722 data tidak menyertakan ulasan. Setelah tahap pre-processing, diperoleh 3850 data observasi yang dapat digunakan dalam proses perhitungan. Pada tahap pengolahan data terdapat operator yang digunakan untuk menyelesaikan masalah ketidakseimbangan kelas yaitu Teknik *Synthetic Minority Oversampling Technique* (SMOTE Upsampling). Teknik tersebut sangat efektif digunakan dalam memodifikasi kumpulan data yang tidak seimbang dengan membuat kelas minoritas sintetik baru untuk meningkatkan efisiensi [23]. Selain itu operator yang digunakan dalam implementasi prosedur klasifikasi adalah total 10 kali (*10 fold*) *cross-validation* dan tipe *automatic sampling*. Dalam Operator *cross-validation* terdapat algoritma *Naive Bayes*, *Decision Tree*, *Support Vector Machine*, dimana masing-masing algoritma dihubungkan dengan operator *Apply Model*, dan *Performance* (*binomial classification*) dengan pengaturan parameter untuk *Accuracy*, *Precision*, *Recall*, *Area Under Curve* (AUC). Adapun, alur pengolahan data di aplikasi Rapidminer dapat dilihat pada Gambar 3 berikut ini.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 3, Desember 2022 Page: 1634−1646
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i3.2766



Gambar 3. Alur Pengolahan Data di Aplikasi Rapidminer

Gambar 3 merupakan alur pengolahan data di aplikasi Rapidminer yang terbagi menjadi kelima bagian yaitu : pertama, alur untuk mengklasifikasi data ulasan dengan luaran visualisasi *wordcloud* hasil klasifikasi data ulasan sesuai kelas negatif dan positif; kedua, alur untuk evaluasi performa algoritma NBC, DT, dan SVM menggunakan operator SMOTE UPsampling dengan pembagian data latih 20% dan data latih 80%; ketiga, alur untuk evaluasi performa algoritma NBC, DT, dan SVM tanpa menggunakan operator SMOTE UPsampling dengan pembagian data latih 20% dan data latih 80%; keempat, hasil evaluasi nilai *Receiver Operating Characteristic* (ROC); kelima, hasil perbandingan performa algoritma menggunakan SMOTE Upsampling dan tanpa menggunakan operator SMOTE UPsampling berdasarkan nilai uji T (T-Test). Dengan demikian, dapat diperoleh informasi tentang algoritma dengan performa terbaik dalam mengolah data teks (data ulasan) untuk menghasilkan rekomendasi optimalisasi produk dan layanan destinasi wisata Candi Borobudur, Indonesia. Dalam mengoptimalkan kinerja produk dan pelayanan di destinasi wisata Candi Borobudur, jumlah kata pada ulasan yang berisi opini positif dapat dijadikan pedoman untuk mengadvokasi permasalahan atraksi, aksesibilitas, amenitas dan akomodasi dengan word count tertinggi. muncul di ulasan. Seperti kata-kata terkait objek wisata yaitu “tempat, borobudur, matahari terbit, kunjungan, candi”, menunjukkan perlunya mengoptimalkan dan menjaga pengelolaan tempat wisata. Sebaliknya, kata-kata dengan emosi negatif dalam ulasan dapat digunakan sebagai bentuk ulasan untuk memperbaiki sistem pengelolaan destinasi sehingga kepuasan pengunjung dapat dipertahankan dan ditingkatkan. Studi ini juga menunjukkan 50 kata yang paling sering muncul di setiap penilaian, seperti yang ditunjukkan pada gambar di bawah ini.



Gambar 4. *Plot Heatmap* 50 Kata yang Terdeteksi dalam Ulasan di *Website Tripadvisor*

Building of Informatics, Technology and Science (BITS)
Volume 4, No 3, Desember 2022 Page: 1634−1646
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i3.2766

Gambar 4 menunjukkan bahwa hasil klasifikasi berupa diagram heatmap, 50 kata paling sering muncul pada pengunjung. Penyortiran berdasarkan 10 kata diharapkan dapat dimasukkan dalam ulasan perasaan positif dan negatif dan paling umum dalam ulasan Candi Borobudur oleh pengguna situs Tripadvisor memberikan kata-kata berikut: kata "candi" ditemukan sebanyak6100 kata dari 2673 ulasan; kata "kunjungan" ditemukan sebanyak 2778 kata dari 1780 ulasan; kata "Borobudur" ditemukan sebanyak 2655 kata dari 1513 ulasan; kata "matahari terbit" ditemukan sebanyak 2435 kata dari 1456 ulasan; kata "lokasi" ditemukan sebanyak 1660 kata dari 1218 ulasan; kata "waktu" ditemukan sebanyak 1573 kata dari 1128 ulasan; kata "panduan" ditemukan sebanyak 1372 kata dari 964 ulasan; kata "tur" ditemukan sebanyak 1356 kata dari 900 ulasan; kata "ambil" ditemukan sebanyak 1.300 kata dari 957 ulasan; kata "orang" ditemukan sebanyak 1070 kata dari 812 ulasan. Hal ini menunjukkan bahwa pemandangan merupakan aspek dominan yang dikomentari pengunjung Candi Borobudur. Selain itu, kata "orang" menunjukkan ekspektasi tentang layanan yang identik dengan keramahan masyarakat lokal, pemandu wisata hingga pengelola destinasi yang memberikan kesan baik. Kajian pengelolaan destinasi wisata menunjukkan bahwa keramahan merupakan salah satu nilai terpenting untuk meningkatkan kepuasan pengunjung [24]. Disisi lain, Hadi & Widyaningsih menunjukkan bahwa daya tarik suatu destinasi dapat berupa nilai-nilai *sapta pesona* yakni keamanan, ketertiban, kebersihan, kesejukan, keindahan, keramahan, dan ketenangan [24]. Hal ini menunjukkan bahwa ulasan pengguna *website Tripadvisor* tentang Candi Borobudur tidak hanya berhubungan dengan atraksi, aksesiblitas, amenitas, dan *ancillary* melainkan juga *sapta pesona*.

Berdasarkan hasil penelitian ini, terdapat beberapa rekomendasi bagi pengelola destinasi wisata Candi Borobudur untuk mengoptimalkan produk dan pelayanan di destinasi wisata sehingga dapat meningkatkan kepuasan wisatawan. Mengingat kata candi, borobudur dan tempat, yang paling sering terkesan memberikan kesan tempat wisata, mengingat intensitasnya maka keaslian candi harus dijaga agar tidak mengalami kehilangan nilai warisan budaya akibat massa. turis. kegiatan kunjungan Selain itu, aspek silahturahmi harus tetap dipertahankan sebagai daya tarik dalam kaitannya dengan budaya masyarakat setempat dalam berinteraksi dan berkomunikasi. Mengingat kata "orang" ditemukan pada 1.070 dari 812 ulasan, maka memperhatikan aspek-aspek yang berkaitan dengan pelayanan prima menjadi keharusan untuk menjaga dan meningkatkan kepuasan pengunjung. Walaupun hasil pengolahan bahan penelitian ini menunjukkan bahwa perasaan positif lebih dominan daripada perasaan negatif, namun citra wisata Candi Borobudur harus dilestarikan agar menjadi daya tarik dan simbol identitas bangsa Indonesia.

3.2 Evaluasi Performa Algoritma NBC, DT dan SVM

Berdasarkan hasil pengolahan 3850 data teks menggunakan operator *cross validation* untuk mengevaluasi performa algoritma Naïve Bayes dalam mengklasifikasi kata ke dalam kelas positif dan kelas negatif di aplikasi Rapidminer, dapat diketahui adanya perbedaan nilai *accuracy*, *recall*, *precision*, serta *Area Under Curve* (AUC) sebelum dan setelah menggunakan operator SMOTE Upsampling. Selain itu, hasil pengujian menggunakan operator *Receiver Operating Characteristic* (ROC) dan Uji T (T-Test) menunjukkan adanya algoritma dengan performa yang dominan dan relevan untuk digunakan sebagai model pengujian dataset (ulasan pengunjung) Candi Borobudur. Beberapa studi terdahulu menunjukkan bahwa algoritma NBC, DT, dan SVM menunjukkan performa yang baik dalam proses klasifikasi [25]–[27]. Lebih jauh, Antinasari et al., menjelaskan bahwa nilai akurasi merupakan tingkat kedekatan antara nilai yang diprediksi dengan nilai aktual, sedangkan nilai presisi ialah tingkat ketepatan prediksi sistem dengan menghitung prediksi benar dari total data yang diprediksi sistem termasuk prediksi salah, sedangkan nilai recall merupakan tingkat keberhasilan mengenali suatu kelas yang harus dikenali [28]. Hal ini menunjukkan bahwa nilai *confusion matrix* sangat penting untuk dianalisis secara komprehensif dalam evaluasi performa masing-masing algoritma. Dengan demikian dapat diketahui bahwa algoritma dengan nilai akurasi, presisi dan *recall* yang lebih baik dapat digunakan sebagai model untuk pengolahan dataset dalam proses analisis sentimen. Pada gambar 4 dapat dilihat nilai confusion matrix algoritma NBC tanpa operator SMOTE Upsampling berdasarkan nilai akurasi, presisi dan *recall*.

accuracy: 95.80% +/- 0.83% (micro average: 95.80%)

	true Negatif	true Positif	class precision
pred. Negatif	131	212	38.19%
pred. Positif	47	5775	99.19%
class recall	73.60%	96.46%	

precision: 99.19% +/- 0.38% (micro average: 99.19%) (positive class: Positif)

	true Negatif	true Positif	class precision
pred. Negatif	131	212	38.19%
pred. Positif	47	5775	99.19%
class recall	73.60%	96.46%	

recall: 96.46% +/- 0.81% (micro average: 96.46%) (positive class: Positif)

	true Negatif	true Positif	class precision
pred. Negatif	131	212	38.19%
pred. Positif	47	5775	99.19%
class recall	73.60%	96.46%	

Gambar 5. *Confusion Matrix* NBC Tanpa Operator SMOTE Upsampling

Building of Informatics, Technology and Science (BITS)
Volume 4, No 3, Desember 2022 Page: 1634−1646
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i3.2766

Gambar 5 menunjukkan nilai confusion matrix dari algoritma NBC tanpa menggunakan operator SMOTE Upsampling, dimana nilai akurasi sebesar 95.80%, nilai pesisi sebesar 99%, dan nilai recall sebesar 96%. Secara spesifik dapat diketahui bahwa hasil klasifikasi True Negatif (TN) dengan Pred.Negatif (PN) sebesar 131, True Positif (TP) dengan Pred. Negatif (PN) sebesar 212. Sedangkan, hasil klasifikasi True Negatif (TN) dengan Pred.Positif (PP) sebesar 47, True Positif (TP) dengan Prediksi Positif (PP) sebesar 5775. Berbeda halnya dengan nilai *confusion matrix* algoritma DT tanpa operator SMOTE Upsampling, pada gambar berikut.

accuracy: 97.55% +/- 0.33% (micro average: 97.55%)

	true Negatif	true Positif	class precision
pred. Negatif	33	6	84.62%
pred. Positif	145	5981	97.63%
class recall	18.54%	99.90%	

precision: 97.63% +/- 0.31% (micro average: 97.63%) (positive class: Positif)

	true Negatif	true Positif	class precision
pred. Negatif	33	6	84.62%
pred. Positif	145	5981	97.63%
class recall	18.54%	99.90%	

recall: 99.90% +/- 0.16% (micro average: 99.90%) (positive class: Positif)

	true Negatif	true Positif	class precision
pred. Negatif	33	6	84.62%
pred. Positif	145	5981	97.63%
class recall	18.54%	99.90%	

Gambar 6. *Confusion Matrix* DT Tanpa Operator SMOTE Upsampling

Gambar 6 menunjukkan nilai confusion matrix dari algoritma DT tanpa menggunakan operator SMOTE Upsampling, dimana nilai akurasi sebesar 97.55%, nilai pesisi sebesar 97.63%, dan nilai recall sebesar 99.90%. Secara spesifik dapat diketahui bahwa hasil klasifikasi True Negatif (TN) dengan Pred.Negatif (PN) sebesar 33, True Positif (TP) dengan Pred. Negatif (PN) sebesar 6. Sedangkan, hasil klasifikasi True Negatif (TN) dengan Pred.Positif (PP) sebesar 145, True Positif (TP) dengan Prediksi Positif (PP) sebesar 5981. Berbeda halnya dengan nilai *confusion matrix* algoritma SVM tanpa operator SMOTE Upsampling, pada gambar berikut.

accuracy: 98.73% +/- 0.54% (micro average: 98.73%)

	true Negatif	true Positif	class precision
pred. Negatif	100	0	100.00%
pred. Positif	78	5987	98.71%
class recall	56.18%	100.00%	

precision: 98.72% +/- 0.54% (micro average: 98.71%) (positive class: Positif)

	true Negatif	true Positif	class precision
pred. Negatif	100	0	100.00%
pred. Positif	78	5987	98.71%
class recall	56.18%	100.00%	

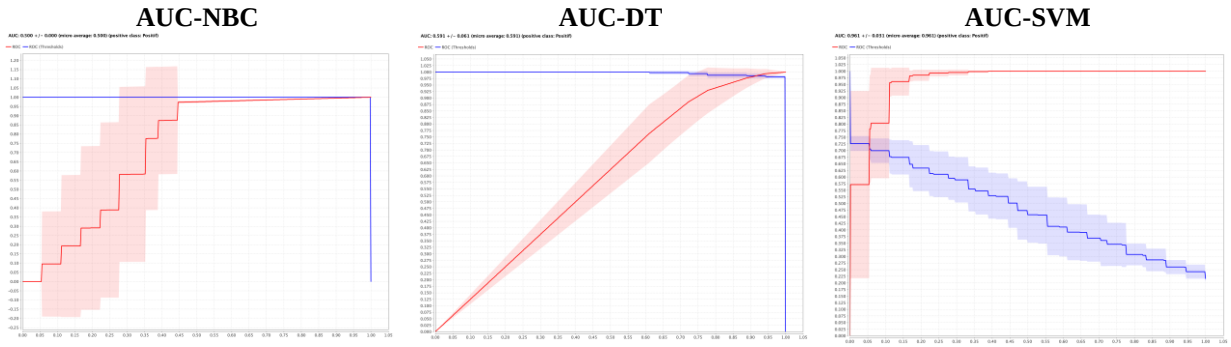
recall: 100.00% +/- 0.00% (micro average: 100.00%) (positive class: Positif)

	true Negatif	true Positif	class precision
pred. Negatif	100	0	100.00%
pred. Positif	78	5987	98.71%
class recall	56.18%	100.00%	

Gambar 7. *Confusion Matrix* SVM Tanpa Operator SMOTE Upsampling

Gambar 7 menunjukkan nilai confusion matrix dari algoritma SVM tanpa menggunakan operator SMOTE Upsampling, dimana nilai akurasi sebesar 98.73%, nilai pesisi sebesar 98.72%, dan nilai recall sebesar 100%. Secara spesifik dapat diketahui bahwa hasil klasifikasi True Negatif (TN) dengan Pred.Negatif (PN) sebesar 100, True Positif (TP) dengan Pred. Negatif (PN) ialah 0 (nol). Sedangkan, hasil klasifikasi True Negatif (TN) dengan Pred.Positif (PP) sebesar 78, True Positif (TP) dengan Prediksi Positif (PP) sebesar 5987. Adapun nilai *Area Under Curve* (AUC) masing-masing algoritma menunjukkan performa yang berbeda sebagaimana gambar berikut.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 3, Desember 2022 Page: 1634−1646
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i3.2766



Gambar 8. Area Under Curve (AUC) algoritma NBC, DT, dan SVM tanpa Operator SMOTE Upsampling

Gambar 8 merupakan nilai Area Under Curve algoritma NBC, DT, dan SVM tanpa SMOTE Upsampling dimana nilai AUC algoritma NBC sebesar 0,50 atau 50%. Selain itu, nilai AUC algoritma DT sebesar 0,591 (59,1%) dan nilai AUC algoritma SVM sebesar 0,961 (96,1%). Hal ini menunjukkan bahwa tanpa menggunakan operator SMOTE Upsampling untuk menyeimbangkan data, performa SVM jauh lebih baik dibandingkan dengan algoritma NBC dan DT.

A	B	C	D
	0.981 +/- 0.003	0.944 +/- 0.007	0.994 +/- 0.002
0.981 +/- 0.003		0.000	0.000
0.944 +/- 0.007			0.000
0.994 +/- 0.002			

Gambar 9. Hasil Uji T algoritma NBC, DT, dan SVM tanpa Operator SMOTE Upsampling

Gambar 9 menunjukkan hasil uji T pada algoritma NBC, DT, dan SVM tanpa operator SMOTE Upsampling yang menunjukkan bahwa algoritma SVM lebih dominan dibandingkan dengan algoritma lainnya, dimana hasil uji T algoritma SVM sebesar 0.994 dibandingkan dengan algoritma NBC sebesar 0.981 dan algoritma DT sebesar 0.944 Dengan demikian dapat disimpulkan bahwa alogirtma SVM lebih dominan dibandingkan algoritma NBC dan DT, meskipun tanpa menggunakan operator SMOTE Upsampling. Selanjutnya, perlu dianalisis secara komprehensif perubahan nilai *Area Under Curve* (ROC) dan nilai hasil uji T (T-test) pada algoritma NBC, DT, dan SVM dengan menggunakan operator SMOTE Upsampling, sebagaimana gambar berikut.

accuracy: 98.08% +/- 0.29% (micro average: 98.08%)

	true Negatif	true Positif	class precision
pred. Negatif	4013	192	95.43%
pred. Positif	0	5795	100.00%
class recall	100.00%	96.79%	

precision: 100.00% +/- 0.00% (micro average: 100.00%) (positive class: Positif)

	true Negatif	true Positif	class precision
pred. Negatif	4013	192	95.43%
pred. Positif	0	5795	100.00%
class recall	100.00%	96.79%	

recall: 96.79% +/- 0.48% (micro average: 96.79%) (positive class: Positif)

	true Negatif	true Positif	class precision
pred. Negatif	4013	192	95.43%
pred. Positif	0	5795	100.00%
class recall	100.00%	96.79%	

Gambar 10. *Confusion Matrix* NBC dengan SMOTE Upsampling

Gambar 10 menunjukkan nilai confusion matrix dari algoritma NBC dengan menggunakan operator SMOTE Upsampling, dimana nilai akurasi sebesar 98.08%, nilai pesisi sebesar 100%, dan nilai recall sebesar 96.79%. Secara spesifik dapat diketahui bahwa hasil klasifikasi True Negatif (TN) dengan Pred.Negatif (PN) sebesar 4013, True Positif (TP) dengan Pred. Negatif (PN) sebesar 192. Sedangkan, hasil klasifikasi True Negatif (TN) dengan Pred.Positif (PP) sebesar 0 (nol), True Positif (TP) dengan Prediksi Positif (PP) sebesar 5795. Berbeda halnya dengan nilai *confusion matrix* algoritma DT tanpa operator SMOTE Upsampling, pada gambar berikut.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 3, Desember 2022 Page: 1634−1646
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i3.2766

accuracy: 94.40% +/- 0.69% (micro average: 94.40%)

	true Negatif	true Positif	class precision
pred. Negatif	3485	32	99.09%
pred. Positif	528	5955	91.86%
class recall	86.84%	99.47%	

precision: 91.86% +/- 0.95% (micro average: 91.86%) (positive class: Positif)

	true Negatif	true Positif	class precision
pred. Negatif	3485	32	99.09%
pred. Positif	528	5955	91.86%
class recall	86.84%	99.47%	

recall: 99.47% +/- 0.22% (micro average: 99.47%) (positive class: Positif)

	true Negatif	true Positif	class precision
pred. Negatif	3485	32	99.09%
pred. Positif	528	5955	91.86%
class recall	86.84%	99.47%	

Gambar 11. *Confusion Matrix* DT dengan operator SMOTE Upsampling

Gambar 11 menunjukkan nilai confusion matrix dari algoritma DT tanpa menggunakan operator SMOTE Upsampling, dimana nilai akurasi sebesar 94.40%, nilai pesisi sebesar 91.86%, dan nilai recall sebesar 99.47%. Secara spesifik dapat diketahui bahwa hasil klasifikasi True Negatif (TN) dengan Pred.Negatif (PN) sebesar 3485, True Positif (TP) dengan Pred. Negatif (PN) sebesar 32. Sedangkan, hasil klasifikasi True Negatif (TN) dengan Pred.Positif (PP) sebesar 528, True Positif (TP) dengan Prediksi Positif (PP) sebesar 5955. Berbeda halnya dengan nilai *confusion matrix* algoritma SVM menggunakan operator SMOTE Upsampling, pada gambar berikut.

accuracy: 99.41% +/- 0.22% (micro average: 99.41%)

	true Negatif	true Positif	class precision
pred. Negatif	4013	59	98.55%
pred. Positif	0	5928	100.00%
class recall	100.00%	99.01%	

precision: 100.00% +/- 0.00% (micro average: 100.00%) (positive class: Positif)

	true Negatif	true Positif	class precision
pred. Negatif	4013	59	98.55%
pred. Positif	0	5928	100.00%
class recall	100.00%	99.01%	

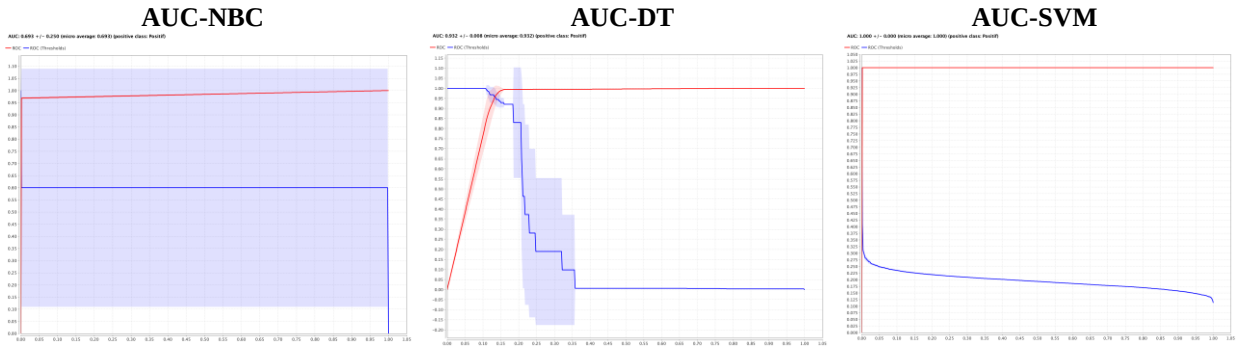
recall: 99.01% +/- 0.37% (micro average: 99.01%) (positive class: Positif)

	true Negatif	true Positif	class precision
pred. Negatif	4013	59	98.55%
pred. Positif	0	5928	100.00%
class recall	100.00%	99.01%	

Gambar 12. *Confusion Matrix* SVM dengan operator SMOTE Upsampling

Gambar 12 menunjukkan nilai confusion matrix dari algoritma SVM menggunakan operator SMOTE Upsampling, dimana nilai akurasi sebesar 99.41%, nilai pesisi sebesar 100%, dan nilai recall sebesar 99.01%. Secara spesifik dapat diketahui bahwa hasil klasifikasi True Negatif (TN) dengan Pred.Negatif (PN) sebesar 4013, True Positif (TP) dengan Pred. Negatif (PN) sebesar 59. Sedangkan, hasil klasifikasi True Negatif (TN) dengan Pred.Positif (PP) ialah 0 (nol), True Positif (TP) dengan Prediksi Positif (PP) sebanyak 5928. Adapun nilai *Area Under Curve* masing-masing algoritma menggunakan operator SMOTE Upsampling menunjukkan performa yang berbeda sebagaimana gambar berikut.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 3, Desember 2022 Page: 1634−1646
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i3.2766



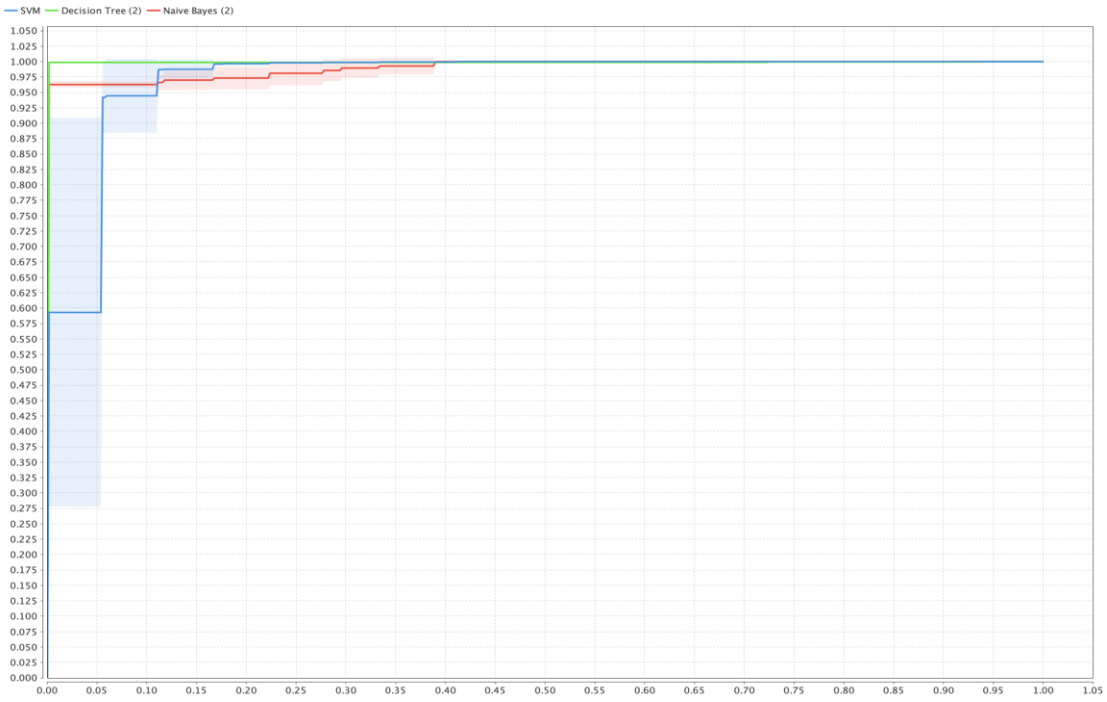
Gambar 13. Area Under Curve (AUC) algoritma NBC, DT, dan SVM menggunakan Operator SMOTE Upsampling

Gambar 13 merupakan nilai Area Under Curve algoritma NBC, DT, dan SVM tanpa SMOTE Upsampling dimana nilai AUC algoritma NBC sebesar 0,693 (69,3%). Selain itu, nilai AUC algoritma DT sebesar 0,932 (93,2%) dan nilai AUC algoritma SVM sebesar 1,00 (100%). Perubahan nilai *Area Under Curve* (AUC) sebelum dan setelah menggunakan operator SMOTE Upsampling menunjukkan adanya performa yang baik dari algoritma yang digunakan. Kadafi menunjukkan bahwa klasifikasi performa *Area Under Curve* (AUC) pada nilai 0,50-0,60 (gagal), 0,60-0,70 (rendah), 0,70-0,80 (adil/sama), 0,80-0,90 (baik), 0,90-1,00 (paling baik) [29]. Dengan demikian dapat diketahui bahwa SVM menunjukkan performa dengan hasil paling baik berdasarkan nilai AUC. Selanjutnya, dapat dilihat hasil uji T (T-test) dari masing-masing algoritma pada gambar berikut.

A	B	C	D
	0.981 +/- 0.003	0.944 +/- 0.007	0.994 +/- 0.002
0.981 +/- 0.003		0.000	0.000
0.944 +/- 0.007			0.000
0.994 +/- 0.002			

Gambar 14. Hasil Uji T Algoritma NBC, DT, dan SVM dengan Operator SMOTE Upsampling

Gambar 14 menunjukkan hasil Uji T (T-test) pada algoritma NBC, DT, dan SVM yang menunjukkan bahwa algoritma SVM lebih dominan dibandingkan dengan algoritma lainnya, dimana hasil uji T algoritma SVM sebesar 0,994 dibandingkan dengan algoritma NBC sebesar 0,981 dan algoritma DT sebesar 0,944. Hal ini menunjukkan bahwa SVM lebih dominan dibandingkan dengan algoritma NBC dan DT dalam pengolahan dataset pengunjung Candi Borobudur. Meskipun demikian, hasil evaluasi algoritma berdasarkan *Receiver Operating Characteristic* (ROC) menunjukkan bahwa algoritma DT juga menunjukkan performa yang baik, sebagaimana gambar berikut.



Gambar 15. ROC Algoritma NBC, DT, dan SVM

Building of Informatics, Technology and Science (BITS)
Volume 4, No 3, Desember 2022 Page: 1634−1646
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i3.2766

Gambar 15 menunjukkan bahwa nilai *Receiver Operating Characteristic* (ROC) menunjukkan bahwa algoritma DT juga menunjukkan performa yang baik, selain SVM. Hal ini menunjukkan bahwa berdasarkan tahapan *modeling* dalam metode CRISP-DM untuk menganalisis sentimen penunjung Candi Borobudur dapat menggunakan algoritma SVM dan DT dengan operator SMOTE Upsampling untuk memperoleh hasil klasifikasi yang lebih baik. Meskipun demikian, masing-masing pengelola destinasi dapat menetapkan model algoritma yang sesuai dengan kondisi eksisting perusahaan atau organisasi pengelola destiansi Candi Borobudur di Indonesia.

4. KESIMPULAN

Hasil penelitian ini menunjukkan bahwa proses klasifikasi data teks berdasarkan sentimen negatif dan positif menggunakan algoritma *Naïve Bayes Classifier, Decision Tree, dan Support Vector Machine* terhadap data ulasan pengguna *website Tripadvisor* dapat diproses menggunakan aplikasi Rapidminer, dengan membandingkan performa algoritma dengan menggunakan operator SMOTE UPsampling maupun tanpa menggunakan operator SMOTE UPsampling. Berdasarkan hasil evaluasi performa algotirma NBC dapat diketahui adanya perubahan nilai *confusion matrix* pada nilai akurasi dari 98.73% menjadi 95.6%, nilai presisi berubah dari 98.72% menjadi 98.97%, nilai *recall* juga berubah dari 100% menjadi 96.54%. Adapun, nilai Area Under Curve (AUC) juga mengalami perubahan dari 0,500 (50%) menjadi 0,693 (69,35%). Selain itu, hasil evaluasi performa algoritma DT menunjukkan perubahan nilai *confusion matrix* pada nilai akurasi dari 97.55% menjadi 94.40%, nilai presisi meningkat dari 97.63% menjadi 91.86%, nilai recall juga berubah dari 99.90% menjadi 99.47%. Adapun, nilai Area Under Curve (AUC) juga mengalami perubahan dari 0,591 (59,1%) menjadi 0,932 (93,2%). Adapun, hasil evaluasi performa algoritma SVM menunjukkan perubahan nilai *confusion matrix* pada nilai akurasi dari 98.73% menjadi 99.41%, nilai presisi berubah dari 98.72% menjadi 100%, nilai recall juga berubah dari 100% menjadi 99.01%. Adapun, nilai Area Under Curve (AUC) juga mengalami perubahan dari 0,961 (96,1%) menjadi 1,00 (100%). Selain itu, hasil uji T menunjukkan bahwa algoritma SVM lebih dominan dibandingkan dengan algoritma lainnya, dimana nilai uji T algoritma SVM sebesar 0,994 jika dibandingkan dengan nilai uji T algoritma DT sebesar 0,944 dan nilai uji T algoritma NBC sebesar 0,98. Berdasarkan nilai *Receiver Operating Characteristic* (ROC) dapat diketahui bahwa algoritma DT juga menunjukkan performa yang baik, selain SVM. Disisi lain, hasil klasifikasi sepuluh kata yang terdeteksi paling sering muncul di setiap ulasan pengunjung ialah sebagai berikut: *temple, visit, borobudur, sunrise, place, time, guide, tour, take, people*. Hal ini menunjukkan bahwa hal-hal yang berkesan bagi pengunjung di Candi Borobudur lebih dominan pada aspek atraksi. Dengan demikian, rekomendasi bagi pengelola destinasi wisata Candi Borobudur agar mempertahankan dan meningkatkan performa layanan produk dan jasa yang berhubungan dengan otentikasi warisan budaya serta mengimplementasikan nilai-nilai sapta pesona untuk meningkatkan kepuasan pengunjung dan citra destinasi wisata Candi Borobudur sebagai destinasi wisata super prioritas di Indonesia.

REFERENSI

[1] L. K. P. Daulay, F. Boy, N. Nakaromi, P. Prakoso, and U. Ramadhanty, "Transformasi Digital Di Ekowisata Bukit Peramun," *J. Ind. Pariwisata*, vol. 5, no. 1, pp. 99–110, 2022, doi: 10.36441/pariwisata.v5i1.991.

[2] G. Hazmin and A. Wijayanti, "Pendekatan Berbasis Phygital dalam Menjembatani Kesenjangan dalam Transformasi Digital," *Int. J. Community Serv. Learn.*, vol. 6, no. 2, pp. 159–166, 2022, doi: 10.23887/ijcsl.v6i2.48470.

[3] N. N. R. Suasih, P. Y. Wijaya, and I. M. E. K. Yudha, "Key Factors Transformasi Digital UMKM (Pendekatan Analisis Micmac Pada Umkm Di Bali)," *J. Akunt. dan Pajak*, vol. 22, no. 2, pp. 1–7, 2022, doi: http://dx.doi.org/10.29040/jap.v22i2.4014.

[4] S. Asril, "Adaptasi Digital: Upaya Menghidupkan Kembali Roh Museum," *War. Pariwisata*, vol. 20, no. 1, pp. 15–17, 2022, doi: 10.5614/wpar.2022.20.1.04.

[5] M. R. Muttaqin, T. I. Hermanto, and M. A. Sunandar, "Penerapan K-Means Clustering Dan Cross-Industry Standard Process for Data Mining (CRISP-DM) Untuk Mengelompokan Penjualan Kue," *Komputasi J. Ilm. Ilmu Komput. dan Mat.*, vol. 19, no. 1, pp. 38–53, 2022, [Online]. Available: https://journal.unpak.ac.id/index.php/komputasi

[6] S. E. A. Felix *et al.*, "A Data Mining-based Cross-Industry Process for Predicting Major Bleeding in Mechanical Circulatory Support," *Eur. Hear. J. - Digit. Heal.*, vol. 2, no. 4, pp. 635–642, 2021, doi: 10.1093/ehjdh/ztab082.

[7] H. N. Prabowo, R. Setyadi, and W. A. Prabowo, "Application of Data Mining for Clustering of Foreign Tourist Visits Based on Arrival Entrance," *Sinkron*, vol. 7, no. 1, pp. 49–58, 2022, doi: 10.33395/sinkron.v7i1.11217.

[8] H. J. Christanto and Y. A. Singgalen, "Sentiment Analysis on Customer Perception towards Products and Services of Restaurant in Labuan Bajo," *J. Inf. Syst. Informatics*, vol. 4, no. 3, pp. 511–523, 2022, doi: 10.51519/journalisi.v4i3.276.

[9] Y. A. Singgalen, "Sentiment Analysis on Customer Perception towards Products and Services of Restaurant in Labuan Bajo," *J. Inf. Syst. Informatics*, vol. 4, no. 3, pp. 511–523, 2022, doi: 10.51519/journalisi.v4i3.276.

[10] A. Wicaksono, N. Khakhim, and N. M. Farda, "Variasi Sentimen Pantai Wisata dari Tweet Berbahasa Indonesia Studi Kasus : Pantai Wisata Di Desa Parangtritis , Kabupaten Bantul," *J. Kepariwisataan, Hosp. dan Perjalanan*, vol. 6, no. 1, pp. 1–15, 2022, doi: 10.34013/jk.v6i1.326.

[11] Y. M. W. Wahyu, A. R. Berto, and E. Murwani, "Analisis Sentimen Jaringan Pesan Kolom Komentar Video Wonderful Indonesia 2022 Jagad Jawi yang Dipengaruhi Budaya," *Avant Garde J. Ilmu Komun.*, vol. 10, no. 2, pp. 201–216, 2022.

[12] N. A. Rahma, Garno, and N. Sulistiyowati, "Analisis Sentimen Tempat Wisata di Jakarta Pasca Covid-19 dengan Algoritma Naive Bayes," *J. Pendidik. dan Konseling*, vol. 4, no. 6, pp. 5894–5908, 2022.

[13] A. A. Arifiyanti, M. F. Pandji, and B. Utomo, "Analisis Sentimen Ulasan Pengunjung Objek Wisata Gunung Bromo pada Situs Tripadvisor," *Explor. J. Sist. Inf. dan Telemat.*, vol. 13, no. 1, p. 32, 2022, doi: 10.36448/jsit.v13i1.2539.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 3, Desember 2022 Page: 1634−1646
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i3.2766

[14] Y. T. Pratama, F. A. Bachtiar, and N. Y. Setiawan, "Analisis Sentimen Opini Pelanggan Terhadap Aspek Pariwisata Pantai Malang Selatan Menggunakan TF-IDF dan Support Vector Machine," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. 12, pp. 6244–6252, 2018.
[15] R. Azmatul 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.
[16] Hartono, O. S. Sitompul, Tulus, and E. B. Nababan, "Biased Support Vector Machine and Weighted-SMOTE in Handling Class Imbalance Problem," *Int. J. Adv. Intell. Informatics*, vol. 4, no. 1, pp. 21–27, 2018, doi: 10.26555/ijain.v4i1.146.
[17] 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.
[18] 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.
[19] R. Puspita and A. Widodo, "Perbandingan Metode KNN, Decision Tree, dan Naïve Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS," *J. Inform. Univ. Pamulang*, vol. 5, no. 4, pp. 646–654, 2021, doi: 10.32493/informatika.v5i4.7622.
[20] D. N. Fitriana and Y. Sibaroni, "Sentiment Analysis on KAI Twitter Post Using Multiclass Support Vector Machine (SVM)," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 5, pp. 846–853, 2020, doi: 10.29207/resti.v4i5.2231.
[21] A. Karim, "Perbandingan Prediksi Kemiskinan di Indonesia Menggunakan Support Vector Machine (SVM) dengan Regresi Linear," *J. Sains Mat. dan Stat.*, vol. 6, no. 1, pp. 107–112, 2020, doi: 10.24014/jsms.v6i1.9259.
[22] E. A. Nida, "Analisis Kinerja Algoritma Support Vector Machine (SVM) Guna Pengambilan Keputusan Beli/Jual Pada Saham PT Elnusa Tbk. (ELSA)," *J. Transform.*, vol. 17, no. 2, pp. 160–170, 2020, doi: 10.26623/transformatika.v17i2.1649.
[23] C. Cahyaningtyas, Y. Nataliani, and I. R. Widiasari, "Analisis Sentimen Pada Rating Aplikasi Shopee Menggunakan Metode Decision Tree Berbasis SMOTE," *Aiti*, vol. 18, no. 2, pp. 173–184, 2021, doi: 10.24246/aiti.v18i2.173-184.
[24] W. Hadi and H. Widyaningsih, "Implementasi Penerapan Sapta Pesona Wisata Terhadap Kunjungan Wisatawan Di Desa Sambirejo Kecamatan Prambanan Kabupaten Sleman Daerah Istimewa Yogyakarta Wisnu," *Khasanah Ilmu J. Pariwisata Dan Budaya*, vol. 11, no. 2, pp. 127–136, 2020, doi: 10.31294/khi.v11i2.8862.
[25] T. Mardiana, H. Syahreva, and T. Tuslaela, "Komparasi Metode Klasifikasi Pada Analisis Sentimen Usaha Waralaba Berdasarkan Data Twitter," *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 267–274, 2019, doi: 10.33480/pilar.v15i2.752.
[26] I. M. B. S. Darma, R. S. Perdana, and Indriati, "Penerapan Sentimen Analisis Acara Televisi Pada Twitter Menggunakan Support Vector Machine dan Algoritma Genetika sebagai Metode Seleksi Fitur," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 3, pp. 998–1007, 2018, [Online]. Available: http://j-ptiik.ub.ac.id
[27] J. Ipmawati, Kusrini, and E. Taufiq Luthfi, "Komparasi Teknik Klasifikasi Teks Mining Pada Analisis Sentimen," *Indones. J. Netw. Secur.*, vol. 6, no. 1, pp. 28–36, 2017.
[28] P. Antinasari, R. S. Perdana, and M. A. Fauzi, "Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 12, pp. 1733–1741, 2017, [Online]. Available: http://j-ptiik.ub.ac.id
[29] A. R. Kadafi, "Perbandingan Algoritma Klasifikasi Untuk Penjurusan Siswa SMA," *J. ELTIKOM*, vol. 2, no. 2, pp. 67–77, 2018, doi: 10.31961/eltikom.v2i2.86.
[30] B. S. Prakoso, D. Rosiyadi, H. S. Utama, and D. Aridarma, "Klasifikasi Berita Menggunakan Algoritma Naive Bayes Classifer Dengan Seleksi Fitur Dan Boosting," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 2, pp. 227–232, 2019, doi: 10.29207/resti.v3i2.1042.
[31] F. Zamachsari, G. Vangeran Saragih, Susafa'ati, and W. Gata, "Analisis Sentimen Pemindahan Ibu Kota Negara dengan Feature Selection Algoritma Naive Bayes dan Support Vector Machine," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 1, no. 3, pp. 504–512, 2017, doi: 10.29207/resti.v4i3.1942.
[32] R. Fatmasari, V. M. Ayu, B. Pratama, and W. Gata, "Analisis Sentimen Dalam Pengkategorian Komentar Youtube Terhadap Layanan Akademik dan Non-Akademik Universitas Terbuka Untuk Prediksi Kepuasan," *Build. Informatics, Technol. Sci.*, vol. 4, no. 2, pp. 395–404, 2022, doi: 10.47065/bits.v4i2.1738.