Klasifikasi Sentimen Wisatawan Candi Borobudur pada Situs *TripAdvisor* Menggunakan *Support Vector Machine* dan *K-Nearest Neighbor*

Rahayu Prihatini Saputri, Wiwiek Setya Winahju, dan Kartika Fithriasari
Departemen Statistika, Fakultas Matematika Komputasi dan Sains Data,
Institut Teknologi Sepuluh Nopember (ITS)
e-mail: kartika_f@statistika.its.ac.id

***Abstrak*—Candi Borobudur merupakan salah satu desti-nasi wisata di Indonesia yang telah dikenal hingga dunia internasional dan kini menjadi satu dari sepuluh destinasi prioritas yang ditetapkan oleh Kementerian Pariwisata. Oleh sebab itu pengelola wisata Candi Borobudur perlu memperhatikan berbagai persepsi wisatawan sebagai bagian dari proses evaluasi. Klasifikasi sentimen wisatawan berdasarkan data ulasan yang tersedia di situs *TripAdvisor* dilakukan dengan metode *Support Vector Machine* (SVM) dan *K-Nearest Neighbor* (K-NN), dengan penerapan teknik *N-gram* di kedua metode tersebut. Selain itu digunakan pula metode *Synthetic Minority Oversampling Technique* (SMOTE) untuk menangani kasus data *imbalance*. Hasil yang diperoleh dari penelitian ini adalah SVM *kernel Radial Basis Function* (RBF) dengan penerapan *unigram* merupakan metode terbaik untuk kasus klasifikasi sentimen wisatawan Candi Borobudur. Kinerja klasifikasi yang dihasilkan oleh metode tersebut tergolong sangat baik.**

Kata Kunci*— *K-Nearest Neighbor, N-gram, Sentimen, Synthetic Minority Oversampling Technique, Support Vector Machine

I. PENDAHULUAN

SENTIMEN merupakan salah satu bentuk aplikasi dari *text mining* yang bertujuan untuk memperoleh pendapat atau sentimen pengguna tentang produk dan layanan yang tersedia [1]. Saat ini telah banyak orang yang memanfaatkan *internet* sebagai media penyampaian ulasan atau pendapatnya. Terdapat banyak situs yang menyediakan fasilitas *review* di situsnya, salah satu diantaranya adalah *TripAdvisor*. Adanya berbagai *review* di situs ini sangat bermanfaat bagi wisatawan karena dapat menjadi tambahan informasi mengenai tempat tujuan, transportasi, dan juga akomodasi [2]. Terdapat salah satu destinasi wisata Indonesia yang terdaftar dalam situs tersebut, yakni Candi Borobudur, dimana tempat wisata tersebut telah dikenal hingga dunia internasional. Selain itu, saat ini Candi Borobudur ditetapkan sebagai salah satu dari sepuluh destinasi wisata yang diprioritaskan oleh Kementerian Pariwisata [3]. Hal tersebut membuat pengelola wisata Candi Borobudur perlu memperhatikan persepsi wisatawan, yang dapat diketahui melalui ulasan wisatawan. Dengan mengetahui berbagai persepsi wisatawan, pengelola wisata Candi Borobudur dapat melakukan evaluasi untuk memperbaiki kekurangan yang ada. Namun diperlukan suatu teknik untuk keefektifan proses evaluasi dari kumpulan data ulasan. Analisis sentimen dapat menjadi teknik penyelesaian permasalahan tersebut.

Dari banyaknya metode klasifikasi yang ada, *Support Vector Machine* (SVM) dan *K-Nearest Neighbor* (K-NN) dapat menjadi pilihan untuk analisis sentimen. K-NN merupakan salah satu metode klasifikasi yang sederhana namun dapat menghasilkan kinerja yang baik [4]. Sedangkan SVM merupakan metode yang efektif untuk *text classification* [5]. Penelitian-penelitian terdahulu juga menunjukkan bahwa metode SVM dan K-NN memiliki kinerja yang baik bila diterapkan pada analisis sentimen. Dalam pendekatan *text mining* dikenal *N-gram*, yaitu teknik menggabungkan beberapa kata sebanyak *n* yang saling berdekatan secara bersamaan [6]. Penerapan *N-gram* dapat menghasilkan informasi yang lebih beragam.

Kinerja klasifikasi yang baik dari metode SVM dan K-NN menjadi pertimbangan digunakannya kedua metode ter-sebut untuk klasifikasi sentimen wisatawan Candi Borobudur berdasarkan data ulasan berbahasa Inggris pada situs *TripAdvisor*. Dalam penelitian ini diterapkan teknik *N-gram* di masing-masing metode klasifikasi sehingga dapat diketahui pengaruhnya terhadap kinerja klasifikasi. Penerapan *N-gram* diduga dapat meningkatkan kinerja klasifikasi [7]. Berdasarkan informasi jumlah ulasan wisatawan Candi Borobudur di situs *TripAdvisor*, terdapat ketimpangan antara jumlah ulasan yang baik dan buruk. Ketimpangan tersebut dapat menjadi perkiraan hasil tahap *labelling* sentimen positif dan negatif menggunakan *lexicon* yang akan mengalami ketimpangan (*imbalance*) pula. Oleh sebab itu pada penelitian ini juga akan digunakan metode untuk mengatasi kasus data *imbalance*, yaitu *Synthetic Minority Oversampling Technique* (SMOTE). Kemudian kinerja metode SVM dan K-NN yang disertai penerapan *N-gram* dievaluasi menggunakan nilai *Area Under the Curve* (AUC).

Berdasarkan penjabaran tersebut, penelitian ini bertu-juan untuk mendapatkan hasil klasifikasi sentimen menggu-nakan metode K-NN dan SVM yang disertai penerapan *N-gram*. Kemudian kedua metode klasifikasi tersebut diban-dingkan kinerjanya, sehingga diperoleh metode terbaik untuk mengklasifikasikan sentimen wisatawan Candi Borobudur. Selain itu penelitian ini juga bertujuan untuk mendapatkan kata-kata yang sering muncul dalam sentimen positif dan negatif. Dengan demikian diperoleh informasi tambahan bagi pihak pengelola wisata Candi Borobudur mengenai hal yang harus diperbaiki atau dipertahankan.

II. TINJAUAN PUSTAKA

A. Text Mining

Text mining merupakan proses penggalian informasi dari suatu data berupa teks yang tidak terstruktur [8]. Bebera-pa permasalahan yang dapat ditangani menggunakan *text mining* yakni klasifikasi, *clustering*, *information extraction*, dan *information retrival* [9].

B. Analisis Sentimen

Analisis sentimen bertujuan untuk menganalisis opini, evaluasi, sikap, penilaian, dan emosi seseorang terhadap enti-

tas tertentu [10]. *Supervised learning* merupakan pendekatan yang dapat diterapkan dalam analisis sentimen, yakni membentuk model klasifikasi berdasrkan data berlabel. Pada penelitian ini diterapkan kamus *lexicon* untuk *labelling* sentimen positif dan negatif terhadap data ulasan, dengan menerapkan persamaan (1) [11].

$$\text{skor} = \text{jumlah kata positif} - \text{jumlah kata negatif} \quad (1)$$

Label positif akan diberikan pada data ulasan dengan skor ≥ 0, sedangkan negatif ketika skor < 0. Penelitian ini menggunakan kamus *lexicon* yang disusun oleh Hu & Liu (2004).

C. *Text Preprocessing*

Text preprocessing dilakukan untuk menghasilkan da-ta yang lebih terstruktur sehingga dapat dilanjutkan analisis dengan *text mining*. Tahap *text preprocessing* dalam penelitian ini adalah sebagai berikut.

1. *Case folding*, yakni mengubah seluruh teks menjadi huruf kecil (non-kapital) dan menghilangkan tanda baca [12].
2. *Lemmatization*, merupakan proses pengubahan kata yang memiliki imbuhan menjadi kata dasarnya (*lemma*), namun tetap sesuai dengan kosakata bahasa Inggris [13].
3. *Stopwords removal*, yaitu menghilangkan kosakata yang tidak dibutuhkan untuk analisis.
4. *Tokenizing*, yaitu pemutusan kata pada kalimat berdasrkan kata-kata yang menyusunnya [14].

D. *N-gram*

N-gram dapat diartikan sebagai teknik penggabungan beberapa kata yang saling berdekatan secara bersamaan sebanyak *n* [6]. *N-gram* dengan *n* = 1 disebut *unigram*, *n* = 2 disebut *bigram*, dan *n* = 3 disebut *trigram*. Dengan meningkatkan ukuran *n*, diperoleh informasi yang lebih beragam. Selain itu penerapan *N-gram* dalam kasus klasifikasi diduga dapat meningkatkan kinerja klasifikasi [7].

E. *Term Weighting*

Term weighting merupakan proses yang penting dalam *text mining* karena dapat mengukur nilai suatu kata dalam keseluruhan dokumen [15]. *Term Frequency-Inverse Document Frequency* (TF-IDF) merupakan metode *term weighting* yang sering digunakan karena cara kerjanya yang sederhana dan efektif. TF-IDF dapat menggambarkan tingkat kepentingan sebuah kata terhadap kumpulan dokumen [16]. Persamaan (2) merupakan rumus pembobotan TF-IDF.

$$w_{ij} = tf_{ij} \times idf \text{ , dengan } idf = \log\left(\frac{N}{df_i}\right), \quad (2)$$

keterangan:

w_{ij} = bobot dari kata *j* pada ulasan ke-*i*
N = jumlah seluruh ulasan
tf_{ij} = jumlah munculnya kata *j* pada ulasan *i*
df_i = banyaknya ulasan yang mengandung kata *j*
i = 1, 2, ..., *M*
j = 1, 2, ..., *m*.

F. *K-fold Cross Validation* (*K-fold* CV)

Salah satu metode untuk mempartisi data menjadi *training* dan *testing* adalah *K-fold Cross Validation* (*K-fold* CV). Metode ini banyak diterapkan oleh peneliti karena bias yang terjadi ketika pengambilan sampel dapat berkurang. Nilai *K* yang umum digunakan dalam *machine lear-ning* adalah 10.

G. *Synthetic Minority Oversampling Technique* (SMOTE)

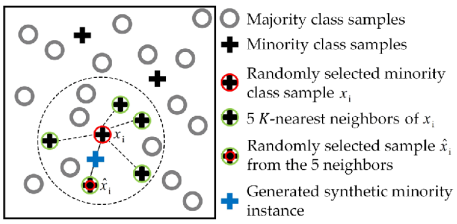
Metode *Synthetic Minority Oversampling Technique* (SMOTE) diusulkan oleh Chawla dkk untuk menangani kasus data tidak seimbang (*imbalance*). Konsep dasar SMOTE adalah membangkitkan data *synthetic* (*oversampling*) berdasarkan tetangga terdekat yang dipilih menggunakan jarak *euclidean*. Hal tersebut mengakibatkan jumlah data di kelas minor menjadi setara dengan kelas mayor [17]. Pembangkitan data *synthetic* dirumuskan dengan persamaan (3) [18].

$$x_{syn} = x_i + (x_{knn} - x_i) \times d, \quad (3)$$

keterangan:

x_{syn} = nilai TF-IDF *synthetic*
x_i = nilai TF-IDF data ke-*i* di kelas minor
x_{knn} = nilai TF-IDF data di kelas minor yang memiliki jarak terdekat dengan x_i
d = bilangan acak antara 0 dan 1

Ilustrasi SMOTE ditunjukkan pada Gambar 1 [19].



Gambar 1. Konsep Dasar SMOTE.

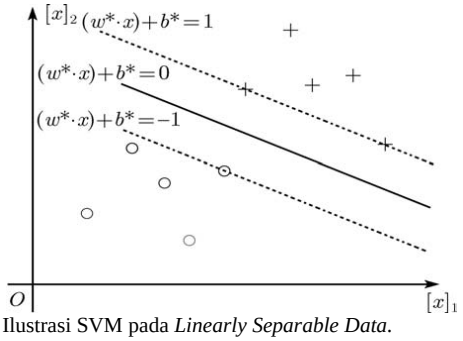
Pada prosedur *K-fold* CV, pembangkitan data *syn-thetic* hanya dilakukan pada data *training* di setiap *fold* untuk menghindari hasil prediksi yang *overoptimistic* [20].

H. *Support Vector Machine* (SVM)

Support Vector Machine (SVM) dikenalkan oleh Vap-nik dkk dan dapat menghasilkan prediksi yang baik pada klasifikasi serta regresi. Klasifikasi pada metode SVM dilakukan berdasrkan fungsi *hyperplane* optimal yang dapat memisahkan dua kelas.

1) SVM pada *Linearly Separable Data*

Misalkan terdapat data *input* $\mathbf{x}_i$ (i = 1, 2, ..., *M*) berdi-mensi *m* dan $y_i \in \{+1, -1\}$ menunjukkan kelas data *input* ke-*i* terpisah secara linier. Maka ilustrasi SVM dua dimensi untuk sepasang variabel X ditampilkan pada Gambar 2 [21].



Gambar 2. Ilustrasi SVM pada *Linearly Separable Data*.

Fungsi pemisah untuk kelas (1) dan (-1) dituliskan dalam bentuk pertidaksamaan (4).

$$\begin{cases} \mathbf{w} \cdot \mathbf{x}_i + b \geq 1; y_i = 1 \\ \mathbf{w} \cdot \mathbf{x}_i + b \leq -1, \end{cases} \quad (4)$$

dengan **w** merupakan vektor bobot dan *b* adalah bias. Kedua kelas tersebut dipisahkan oleh *margin* sebesar $\frac{2}{\|\mathbf{w}\|}$ [22]. *Hyperplane* optimal diperoleh dengan memaksimumkan fungsi objektif $\frac{2}{\|\mathbf{w}\|}$ atau meminimumkan fungsi objektif $\frac{1}{2}\|\mathbf{w}\|^2$, terhadap *constraint* $y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1 \geq 0$. Optimasi tersebut diselesaikan dengan menggunakan formula *Lagrange primal* (L_P) yang diformulasikan pada persamaan (5) [23].

$$L_P(\mathbf{w},b,\boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{M} a_i\{y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1\}, \quad (5)$$

dengan $a_i \geq 0$ merupakan nilai koefisien *Lagrange*. Kemudian persamaan (5) dioptimasi dengan cara meminimumkan L_P terhadap **w** dan *b*. Hasil optimasi tersebut selanjutnya disubstitusi ke dalam persamaan (5) sehingga diperoleh formula *Lagrange dual* (L_D), yakni persamaan (6).

$$L_D = \sum_{i=1}^{M} a_i - \frac{1}{2}\sum_{i,h=1}^{M} a_i a_h y_i y_h \mathbf{x}_i\cdot\mathbf{x}_h . \quad (6)$$

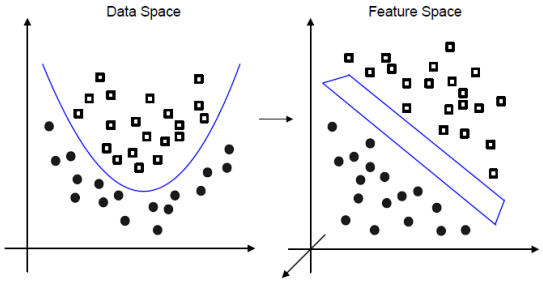
Memaksimumkan L_D terhadap α dengan syarat $\sum_{i=1}^{M} a_i y_i = 0$ dan $a_i \geq 0$ akan menghasilkan nilai α_i. Data *input* dikatakan sebagai *support vector* apabila $\alpha_i > 0$. Sedangkan data *input* dengan nilai $\alpha_i = 0$ disebut *non-support vector*. Setelah nilai α_i diketahui, maka nilai tersebut digunakan untuk memperoleh **w**. Sehingga dihasilkan persamaan (7) untuk klasifikasi dengan metode SVM pada data yang terpisah secara linier.

$$f(\mathbf{x}) = \sum_{i=1}^{sv} a_i y_i \mathbf{x}_i\cdot\mathbf{x} + b . \quad (7)$$

Notasi $\mathbf{x}_i$ adalah *support vector*, nilai *b* diperoleh melalui perhitungan $b = \frac{1}{2}[(\mathbf{w}.\mathbf{x}^*(1)) + (\mathbf{w}.\mathbf{x}^*(-1))]$, dan $i = 1, 2, ..., sv$ (banyaknya *support vector*). Adapun notasi $\mathbf{x}^*(1)$ menunjukkan *support vector* yang berada di kelas +1, sedangkan $\mathbf{x}^*(-1)$ untuk kelas -1 [24].

2) SVM pada *Non-Linearly Separable Data*

Pada kasus riil tidak semua data dapat dipisahkan secara linier [25], sehingga formula SVM dimodifikasi agar dapat menyelesaikan kasus data tidak terpisah secara linier. SVM untuk *non-linearly separable data* diilustrasikan pada Gambar 3 [25].



Gambar 3. Ilustrasi SVM pada *Non Linearly Separable Data*.

Modifikasi yang dilakukan terhadap SVM adalah penambahan fungsi *kernel* yang dituliskan dalam bentuk persamaan (8).

$$K(\mathbf{x}_i,\mathbf{x}_h) = f(\mathbf{x}_i)\cdot f(\mathbf{x}_h) . \quad (8)$$

Adanya fungsi *kernel* mengakibatkan fungsi klasifikasi pada persamaan (7) berubah menjadi persamaan (9) [22].

$$f(\mathbf{x}) = \sum_{i=1}^{sv} a_i y_i K(\mathbf{x},\mathbf{x}_i) + b . \quad (9)$$

Lima jenis *kernel* yang umum digunakan dalam metode SVM adalah *Linear*, *Polynomial*, *Radial Basis Function* (RBF) atau *Gaussian*, *Sigmoid*, dan *Inverse Multiquadric Function* [26].

I. K-Nearest Neighbor (K-NN)

K-Nearest Neighbor (K-NN) dikenal sebagai metode klasifikasi yang populer dalam *data mining* dan statistik karena cara kerjanya yang sederhana namun memiliki kinerja yang baik [27][4]. Pada metode ini, pengklasifikasian dilakukan berdasarkan kategori dari *k* tetangga terdekat antara data *testing* dengan seluruh data *training* [28]. Nilai *k* ditentukan melalui *trial and error*, yakni melakukan uji coba dengan nilai *k* yang berbeda-beda hingga diperoleh *error* yang paling minimum [29]. Selain *error*, dapat pula dipertimbangkan berdasarkan akurasi [30]. Secara umum terdapat empat langkah untuk mendapatkan kategori data *testing* pada metode K-NN, yakni sebagai berikut.

1. Menentukan nilai *k*.
2. Menghitung jarak *euclidean* antara data *testing* dengan setiap data *training*. Jarak *euclidean* dirumuskan pada persamaan (10).

$$d(1,2) = \sqrt{\sum_{j=1}^{m} (x_{1j} - x_{2j})^2} , \quad (10)$$

   dimana notasi $d(1,2)$ adalah jarak *euclidean* antara data *testing* dengan *training*, x_{1j} adalah nilai TF-IDF di data *training*, x_{2j} merupakan nilai TF-IDF data *testing*, dan *m* menunjukkan jumlah *feature* (kata).
3. Mengurutkan jarak *euclidean* dari yang terendah hingga tertinggi, sehingga dapat diketahui *k* tetangga terdekat.
4. Klasifikasi data *testing* berdasarkan kelas yang memiliki jumlah anggota terbanyak di *k* tetangga terdekat.

J. Evaluasi Kinerja Klasifikasi

Kinerja suatu metode klasifikasi dievaluasi menggunakan tiga kriteria yang umum digunakan, yaitu akurasi, sensitivitas (*recall*), dan spesifisitas [31]. Persamaan (11), persamaan (12), dan persamaan (13) merupakan formula untuk memperoleh nilai akurasi, sensitivitas, dan spesifisitas.

$$\text{akurasi} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} , \quad (11)$$

$$\text{sensitivitas} = \frac{\text{TP}}{\text{TP+FN}} , \quad (12)$$

$$\text{spesifisitas} = \frac{\text{TN}}{\text{TN+FP}} . \quad (13)$$

Pada kelas biner, perumusan ketiga unsur evaluasi tersebut didasarkan pada *confusion matrix* yang ditampilkan pada Tabel 1.

Tabel 1.
Confusion Matrix

Kelas Aktual	Kelas Prediksi	
	Negatif	Positif
Negatif	*True Negative* (TN)	*False Positive* (FP)
Positif	*False Negative* (FN)	*True Positive* (TP)

Akurasi didefinisikan sebagai ukuran efektivitas secara keseluruhan dari suatu metode klasifikasi [32]. Namun ketika ditemui kasus data *imbalance*, akurasi diganti dengan *Area Under the Curve* (AUC), yakni kriteria evaluasi yang menggabungkan aspek sensitivitas dan spesifisitas. Nilai AUC dirumuskan pada persamaan (14).

$$\text{AUC} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP+FN}} + \frac{\text{TN}}{\text{TN+FP}}\right). \quad (14)$$

Nilai AUC umumnya berada pada interval 0,5 - 1,0. Keterangan setiap interval nilai AUC dijelaskan pada Tabel 2 [31].

Tabel 2.
Kategori Nilai AUC

Nilai AUC	Keterangan
0,5 – 0,6	*Poor*
0,6 – 0,7	*Fair*
0,7 – 0,8	*Good*
0,8 – 0,9	*Very Good*
0,9 – 1,0	*Excellent*

K. Word Cloud

Visualisasi dalam *text mining* sering disajikan menggunakan *word cloud*. Melalui *word cloud* dapat diperoleh informasi mengenai kata-kata yang sering muncul dalam dokumen. Kata dengan frekuensi kemunculan tinggi ditandai dengan ukuran kata yang besar dalam *word cloud* tersebut [33].

L. Web Scraping

Web scraping adalah teknik pengambilan data atau dokumen semi-terstruktur dari suatu *website* sehingga diperoleh informasi, baik secara keseluruhan maupun sebagian, yang dapat digunakan untuk kepentingan tertentu. Umumnya teknik *web scraping* terdiri dari empat tahap, yakni *create scraping template*, *explore site navigation*, *automate navigation and extraction*, dan *extracted data and package history*.

M. TripAdvisor

TripAdvisor merupakan situs wisata terbesar di dunia yang memiliki lebih dari 730 juta opini dan ulasan dalam si-tusnya. Tersedianya ulasan dalam situs ini dapat membantu wisatawan untuk memutuskan pilihan mengenai tempat tujuan, transportasi, serta akomodasi.

N. Candi Borobudur

Candi Borobudur adalah candi bercorak Buddha terbe-sar di dunia yang terletak di Kabupaten Magelang, Provinsi Jawa Tengah. Destinasi wisata ini sering dikunjungi oleh le-bih dari satu juta wisatawan, baik domestik maupun mancanegara. Sejak tahun 2016, Candi Borobudur ditetapkan menjadi salah satu dari sepuluh destinasi wisata yang diprioritaskan oleh Kementerian Pariwisata.

III. METODOLOGI PENELITIAN

A. Sumber Data

Data yang digunakan dalam penelitian ini berupa kumpulan ulasan atau *review* berbahasa Inggris mengenai wisata Candi Borobudur di situs *TripAdvisor*. Sejak Oktober 2005 – 11 Maret 2019, terdapat 3591 data ulasan yang terkumpul. Data tersebut diperoleh dengan melakukan teknik *web scraping*.

B. Variabel Penelitian

Variabel dalam penelitian ini terdiri dari dua jenis, ya-itu frekuensi kata dan kelas sentimen. Kedua variabel tersebut dijelaskan lebih rinci pada Tabel 3.

Tabel 3.
Variabel Penelitian

Variabel	Keterangan	Skala Data
Y	Kelas sentimen 0 = Positif 1 = Negatif	Nominal
X	Kata (Frekuensi)	Rasio

C. Langkah Analisis

Terdapat beberapa tahap analisis yang diterapkan pada penelitian ini, yaitu sebagai berikut.

1. Mengumpulkan data berupa ulasan berbahasa Inggris wisatawan Candi Borobudur di situs *TripAdvisor*.
2. Melakukan *text preprocessing* yang terdiri dari *case folding*, *lemmatization*, *stopwords removal*, dan *tokenizing*.
3. *Labelling* sentimen positif dan negatif berdasarkan kamus *lexicon* yang disusun oleh Hu & Liu.
4. Membentuk *N-gram* berukuran $n = 1$ (*unigram*), $n = 2$ (*bigram*), dan $n = 3$ (*trigram*).
5. Menganalisis karakteristik data ulasan wisatawan.
6. Menghitung nilai bobot TF-IDF di setiap *N-gram*.
7. Mempartisi data ulasan menjadi *training* dan *testing* dengan 10-*fold Cross Validation* (10-*fold* CV) di setiap *N-gram*.
8. Membangkitkan data *synthetic* (*oversampling*) pada setiap *N-gram* dengan metode SMOTE. *Oversampling* terhadap kelas minor hanya dilakukan di data *training*.
9. Klasifikasi sentimen wisatawan Candi Borobudur untuk setiap *N-gram* menggunakan K-NN.
10. Klasifikasi sentimen wisatawan Candi Borobudur dengan metode SVM di setiap *N-gram*. Tahap-tahap klasifikasi pada metode SVM adalah sebagai berikut.
    a. Menentukan parameter *C* pada *kernel Linear*, serta parameter *C* dan γ untuk *kenel* RBF.
    b. Membuat model SVM berdasarkan jenis *kernel* dan *N-gram* terbaik.
11. Evaluasi kinerja klasifikasi dan pemilihan metode klasifikasi terbaik berdasarkan nilai rata-rata AUC 10-*fold* CV yang tertinggi di data *testing*.
12. Visualisasi sentimen positif dan negatif berdasarkan ukuran *N-gram* terbaik.
13. Interpretasi, menarik kesimpulan, dan saran.

IV. ANALISIS DAN PEMBAHASAN

A. Text Preprocessing

Text preprocessing bertujuan menghilangkan bebera-pa unsur yang tidak dibutuhkan pada tahap analisis *text mi-ning*. Unsur-unsur tersebut dapat berupa tanda baca, angka, dan kata penghubung atau keterangan. *Text preprocessing* da-lam penelitian ini meliputi *case folding*, *lemmatization*, *stopwords removal*, dan *tokenizing*. Hasil *text preprocessing* tersebut diilustrasikan pada Tabel 4.

Tabel 4.
Ilustrasi *Text Preprocessing*

Kalimat Awal	Hasil *Text Preprocessing*
Amazing place, sunrise was great, but it is worth to spend time to see the temple proper-ly with a guide to understand the complexity of the place and the work done.	"*amaze*" "*sunrise*" "*great*" "*worth*" "*spend*" "*proper-ly*" "*guide*" "*understand*" "*complexity*" "*work*"

B. Labelling

Setelah dilakukan *text preprocessing* terhadap seluruh data ulasan, selanjutnya dilakukan *labelling* berdasarkan ka-mus *lexicon* untuk memperoleh label sentimen positif dan ne-

gatif. Label sentimen suatu ulasan dapat diketahui dengan menerapkan persamaan (1), sehingga didapatkan hasil seperti ilustrasi pada Tabel 5.

Tabel 5.
Ilustrasi *Labelling* dengan *Lexicon*

Hasil *Text Preprocessing*	Skor	Label Sentimen
amaze sunrise great worth spend properly guide understand complexity work	4 – 0 = 4	Positif
sunrise crowd largest heritage impress scale masterpiece	2 – 1 = 1	Positif
absolutely overprice entrance public bus accommodation approximate cost visit	0 – 1 = -1	Negatif

Huruf tebal menandakan bahwa suatu kata memiliki makna positif dalam kamus *lexicon*. Sedangkan kata bermakna nega-tif ditandai dengan huruf tebal dan memiliki garis bawah.

Metode lain yang dapat diterapkan untuk *labelling* da-ta yang memiliki dua kategori kelas adalah regresi logistik bi-ner, dimana metode ini sangat dikenal dalam keilmuan statis-tika. Label kelas suatu data dapat diketahui berdasarkan va-riabel prediktor yang berupa kata, serta variabel respon yang tidak lain adalah kategori kelas (0 dan 1).

C. *N-gram*

Penerapan *N-gram* bermanfaat agar informasi yang di-peroleh lebih beragam. Selain itu terdapat dugaan bahwa de-ngan menerapkan *N-gram*, maka hasil klasifikasi yang diper-oleh menjadi lebih baik. Oleh sebab itu dalam penelitian ini diterapkan teknik *N-gram* dengan ukuran n = 1 (*unigram*), n = 2 (*bigram*), dan n = 3 (*trigram*) di metode SVM dan K-NN untuk melihat pengaruhnya terhadap kinerja klasifikasi. Pem-bentukan *feature* pada *unigram*, *bigram*, dan *trigram* diilus-trasikan pada Tabel 6 berdasarkan hasil ilustrasi *text prepro-cessing* pada Tabel 4.

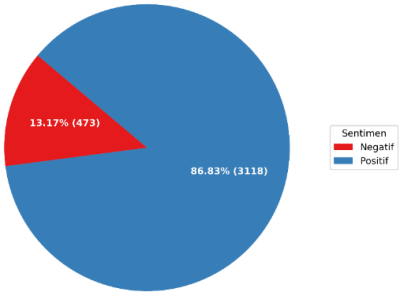
Tabel 6.
Ilustrasi *N-gram*

N-gram	Hasil
Unigram	"*amaze*" "*sunrise*" "*great*" "*worth*" "*spend*" "*properly*" "*guide*" "*understand*" "*complexity*" "*work*"
Bigram	"*amaze sunrise*" "*sunrise great*" "*great worth*" "*worth spend*" "*spend properly*" "*properly guide*" "*guide understand*" "*understand complexity*" "*complexity work*"
Trigram	"*amaze sunrise great*" "*sunrise great worth*" "*great worth spend*" "*worth spend properly*" "*spend properly guide*" "*properly guide understand*" "*guide understand complexity*" "*understand complexity work*"

D. *Karakteristik Ulasan*

Data ulasan berbahasa Inggris wisatawan Candi Boro-budur yang terkumpul melalui *scraping* terhadap *website TripAdvisor* adalah 3591 ulasan. Setelah dilakukan *labelling* berdasarkan *lexicon*, dapat diketahui persepsi wisatawan ter-hadap Candi Borobudur yang ditampilkan pada Gambar 4.

Gambar 4 menunjukkan bahwa lebih banyak wisata-wan yang memiliki persepsi positif terhadap Candi Borobu-dur. Hal ini terlihat dari persentase ulasan bersentimen positif yang mencapai angka 86,83%. Sedangkan ulasan bersenti-men negatif hanya memiliki persentase sebesar 13,17%. Tidak seimbangnya jumlah data di kelas positif dan negatif menjadi indikasi bahwa penelitian ini mengalami kasus *imbalance* sehingga dilakukan pembuatan data *synthetic* (*o-versampling*) pada kelas minor (negatif) menggunakan meto-de SMOTE di kedua metode klasifikasi.



Gambar 4. Perbandingan Label Sentimen Positif dan Negatif.

E. *Klasifikasi Sentimen menggunakan K-Nearest Neighbor (K-NN)*

Kinerja metode K-NN sangat dipengaruhi oleh nilai k. Metode K-NN dengan teknik *unigram* memiliki kinerja pa-ling optimal ketika k = 2, yakni dengan rata-rata AUC bernilai sebesar 0,6139. Untuk *bigram* dan *trigram*, rata-rata AUC yang tertinggi terdapat di k = 17 dan k = 22. Nilai rata-rata AUC pada *bigram* adalah 0,5475, sedangkan pada *trigram* adalah 0,5235. Perbandingan kinerja metode K-NN antara *unigram*, *bigram*, dan *trigram* dengan menggunakan nilai k terbaik ditampilkan pada Tabel 7.

Tabel 7.
Performa Klasifikasi K-NN

Ukuran Evaluasi	*N-gram*		
	Unigram	*Bigram*	*Trigram*
Akurasi	0,41	0,48	0,82
Spesifisitas	0,34	0,46	0,93
Sensitivitas	0,89	0,63	0,11
AUC	0,6139	0,5475	0,5235

Perbandingan kinerja K-NN antara tiga jenis *N-gram* menun-jukkan bahwa *unigram* (n = 1) merupakan jenis *N-gram* ter-baik untuk metode K-NN. Hal tersebut terlihat dari nilai rata-rata AUC pada *unigram* yang lebih tinggi dibandingkan *bi-gram* dan *trigram*, yakni 0,6139. Merujuk pada Tabel 2, dapat dikatakan bahwa metode K-NN dengan hanya menerapkan teknik *unigram* tergolong memiliki performansi yang cukup atau *fair*.

F. *Klasifikasi Sentimen menggunakan Support Vector Machine (SVM)*

SVM dikenal sebagai metode klasifikasi yang memi-liki kinerja baik. Terdapat berbagai jenis *kernel* pada metode SVM untuk mengklasifikasikan data yang terpisah secara li-nier maupun non-linier. Dari lima jenis *kernel* yang ada, pe-nelitian ini hanya menggunakan dua *kernel*, yaitu *Linear* dan *Radial Basis Function* (RBF).

1) SVM *Kernel Linear*

Parameter yang digunakan pada SVM *kernel Linear* a-dalah *cost* (*C*). Pemilihan nilai *C* untuk memprediksi data *tes-ting* dilakukan dengan menggunakan nilai *C* yang berbeda-beda di setiap *N-gram*, kemudian dipilih nilai *C* yang meng-hasilkan rata-rata nilai AUC tertinggi pada data *testing*. Ren-tang nilai *C* yang digunakan dalam penelitian ini adalah 10^{-2} hingga 10^{2}. Tabel 8 adalah perbandingan rata-rata nilai AUC dari setiap nilai *C* untuk *unigram*, *bigram*, dan *trigram*.

Tabel 8.
Perbandingan Nilai *C* terhadap SVM *Kernel Linear*

N-gram	*C*				
	10^{-2}	10^{-1}	10^{0}	10^{1}	10^{2}
Unigram	0,7551	0,8142	0,8143	0,7673	0,7514
Bigram	0,6009	0,6001	0,5745	0,5842	0,5840
Trigram	0,5184	0,5331	0,5335	0,5268	0,5266

Untuk setiap jenis *N-gram*, nilai *C* yang menghasilkan kinerja paling optimal secara berturut-turut adalah 10^0, 10^{-2}, dan 10^0. Dengan demikian dapat diperoleh berbagai ukuran evaluasi, yang ditampilkan pada Tabel 9.

Tabel 9.
Performa Klasifikasi SVM *Kernel Linear*

Ukuran Evaluasi	*N-gram*		
	Unigram	*Bigram*	*Trigram*
Akurasi	0,88	0,77	0,29
Spesifisitas	0,90	0,82	0,20
Sensitivitas	0,73	0,38	0,86
AUC	0,8143	0,6009	0,5335

Di antara *unigram*, *bigram*, dan *trigram*, metode SVM *kernel Linear* akan menghasilkan kinerja paling optimal ketika me-nerapkan *unigram*. Rata-rata nilai AUC yang dihasilkan me-tode ini mampu mencapai 0,8143, sehingga tergolong memi-liki kinerja klasifikasi sangat baik.

2) SVM *Kernel* RBF

Pada *kernel* RBF, terdapat dua parameter yang menentukan kinerja klasifikasi metode SVM, yaitu *C* dan *gamma* (γ). Penentuan nilai terbaik untuk kedua parameter tersebut juga dilakukan melalui *trial and error*. Maka dari itu digunakan nilai $C = 10^{-2}, 10^{-1}, ..., 10^2$ dan $\gamma = 2^{-15}, 2^{-13}, ..., 2^3$ di setiap jenis *N-gram* untuk melihat pengaruhnya terhadap kinerja klasifikasi, sehingga dapat diperoleh nilai kombinasi parameter yang menghasilkan kinerja paling optimum. Rata-rata AUC setiap *N-gram* dari kombinasi nilai *C* dan γ ditampilkan pada Tabel 10.

Tabel 10.
Kombinasi Nilai Parameter Terbaik

N-gram	Parameter		Rata-rata AUC
	C	*Γ*	
Unigram	10^2	2^{-9}	0,8234
Bigram	10^{-1}	2^{-5}	0,6068
Trigram	10^1	2^{-5}	0,5417

Setelah mendapatkan kombinasi nilai *C* dan γ yang terbaik, maka dapat dilakukan evaluasi kinerja klasifikasi de-ngan akurasi, sensitivitas, spesifisitas, dan AUC yang ditam-pilkan pada Tabel 11.

Tabel 11.
Kinerja Klasifikasi SVM *Kernel* RBF

Ukuran Evaluasi	*N-gram*		
	Unigram	*Bigram*	*Trigram*
Akurasi	0,87	0,69	0,31
Spesifisitas	0,89	0,71	0,23
Sensitivitas	0,76	0,50	0,86
AUC	0,8234	0,6068	0,5417

Pada SVM *kernel* RBF, $n = 1$ juga merupakan ukuran *N-gram* yang menghasilkan kinerja klasifikasi paling optimal. Pene-rapan *unigram* pada SVM *kernel* RBF menghasilkan rata-rata nilai AUC sebesar 0,8234. Maka dapat dikatakan bahwa me-tode ini memiliki kinerja klasifikasi yang sangat baik.

3) Model *Support Vector Machine* (SVM)

Perbandingan nilai rata-rata AUC antara *kernel Linear* dan RBF menunjukkan bahwa *kernel* RBF dengan penerapan *unigram* memiliki kinerja klasifikasi yang lebih baik. Dengan parameter $C = 10^2$ dan $\gamma = 2^{-9} \approx 0{,}002$, maka model SVM *kernel* RBF dengan teknik *unigram* dapat ditulis ke dalam persamaan 15.

$$f(\mathbf{x}) = \sum_{i=1}^{2471} a_i y_i K(\mathbf{x},\mathbf{x}_i) - 41{,}1822 . \quad (15)$$

Notasi y_i adalah kelas setiap *support vector* (+1,-1), a_i adalah koefisien *Lagrange*, angka -41,1822 menunjukkan bias (*b*), dan $i = 1, 2, ..., 2471$, merupakan banyaknya *support vector*. Adapun $K(\mathbf{x},\mathbf{x}_i) = exp(-0{,}002 ||\mathbf{x} - \mathbf{x}_i||^2)$ merupakan fungsi *kernel* RBF.

G. Perbandingan Metode Klasifikasi

Untuk memilih metode terbaik yang dapat diterapkan pada kasus klasifikasi sentimen wisatawan Candi Borobudur, maka dilakukan perbandingan antara kinerja metode SVM dengan K-NN berdasarkan akurasi, sensitivitas, spesifisitas, dan AUC di setiap jenis *N-gram*. Perbandingan tersebut ditampilkan pada Tabel 12.

Tabel 12.
Performa Klasifikasi antar Metode

Metode	Ukuran Evaluasi	*N-gram*		
		Unigram	*Bigram*	*Trigram*
SVM (RBF)	Akurasi	0,87	0,69	0,31
	Spesifisitas	0,89	0,71	0,23
	Sensitivitas	0,76	0,50	0,86
	AUC	0,8234	0,6068	0,5417
K-NN	Akurasi	0,41	0,48	0,82
	Spesifisitas	0,34	0,46	0,93
	Sensitivitas	0,89	0,63	0,11
	AUC	0,6139	0,5457	0,5235

Melalui Tabel 12 dapat diketahui bahwa metode terbaik yang dapat digunakan untuk persoalan klasifikasi sentimen wisata-wan Candi Borobudur adalah SVM *kernel* RBF yang disertai teknik *unigram*. Metode tersebut dipilih karena memiliki nilai rata-rata AUC yang paling tinggi di antara metode lainnya, yakni mencapai 0,8234, yang mengindikasikan bahwa meto-de ini memiliki kinerja klasifikasi yang sangat baik. Tabel 12 juga menunjukkan bahwa pada penelitian ini, bertambahnya ukuran *n* pada *N-gram* justru menurunkan nilai AUC. Hal ter-sebut dapat terjadi karena tahap *labelling* dalam penelitian ini hanya berdasarkan satu kata saja. Beberapa perbedaan antara hasil *labelling* dengan *lexicon* dan *N-gram* disajikan pada Ta-bel 13, dimana hasil *labelling* dengan *N-gram* diperoleh dari sudut pandang peneliti.

Tabel 13.
Perbedaan Hasil *Labelling* dengan *Lexicon* dan *N-gram*

No.	*Review*	*Lexicon*	*Bigram*	*Trigram*
1	*wonder world unesco world heritage site*	Positif	Positif	Positif
2	*unbelievable architecture shock visit sunrise sunset bit crow-dy sunset*	Negatif	Positif	Positif
3	*highly recommend pay extra guide interest fact overcrowd*	Positif	Negatif	Negatif

Dari tiga contoh data ulasan yang ditampilkan dalam Tabel 13, terdapat dua data dengan hasil *labelling* yang berbeda antara *lexicon* dan *N-gram*. Perbedaan tersebut dapat berpengaruh pada hasil klasifikasi ketika teknik *unigram*, *bigram*, dan *trigram* diterapkan di suatu metode klasifikasi.

Untuk mengetahui jumlah data yang terklasifikasi de-ngan benar ketika menggunakan metode SVM *kernel* RBF dan teknik *unigram*, maka pada Tabel 14 ditampilkan hasil *confusion matrix*. Ketika menggunakan metode SVM *kernel* RBF dan teknik *unigram*, terdapat 460 data yang salah terklasifikasi (misklasifikasi). 115 dari 473 ulasan bersentimen negatif justru diklasifikasikan menjadi sentimen positif, atau dapat dikatakan persentase kesalahan klasifikasi pada kelas negatif adalah sebesar 24%. Sedangkan untuk kelas positif, per-

sentase misklasifikasi bernilai sebesar 11%, dengan rincian terdapat 345 data ulasan yang diklasifikasikan menjadi kelas negatif.

Tabel 14.
Confusion Matrix SVM *Kernel* RBF (*Unigram*)

Kelas Aktual	Kelas Prediksi		Total
	Positif	Negatif	
Positif	2773	345	3118
Negatif	115	358	473
Total	2888	703	3591

Selanjutnya dilakukan perbandingan kinerja klasifika-si antara data *training* dengan *testing* pada teknik *unigram* menggunakan metode SVM Kernel RBF. Perbandingan ter-sebut ditampilkan pada Tabel 15 dan bertujuan untuk meng-identifikasi apakah model klasifikasi yang terbentuk telah *fit* atau dapat digunakan untuk memprediksi data-data ulasan di waktu yang akan datang**.**

Tabel 15.
Perbandingan Kinerja Klasifikasi Data *Training* dan *Testing* pada SVM-RBF-*Unigram*

Data	Akurasi	AUC
Training	0,95	0,9466
Testing	0,87	0,8234

Pada Tabel 15 terlihat bahwa hasil klasifikasi antara data *training* dan *testing* tidak jauh berbeda. Sehingga dapat dika-takan bahwa model klasifikasi yang terbentuk pada data *trai-ning* tidak mengalami kasus *overfitting* maupun *underfitting*, dan model tersebut dapat digunakan untuk memprediksi sen-timen wisatawan Candi Borobudur ketika muncul data ulasan baru.

H. Word Cloud

Dari tiga jenis *N-gram* yang dibandingkan, diperoleh hasil bahwa *unigram* merupakan jenis *N-gram* terbaik yang dapat diterapkan dalam kasus ini. Dengan demikian diperoleh Gambar 5 dan Gambar 6 yang menunjukkan *word cloud uni-gram* untuk sentimen positif dan negatif.

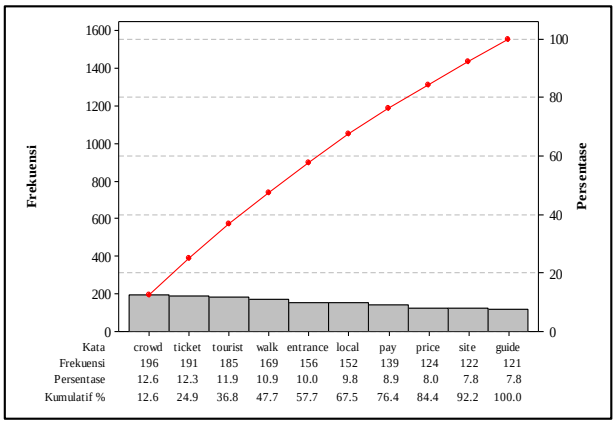
Gambar 5. *Word Cloud Unigram* Positif.

Gambar 6. *Word Cloud Unigram* Negatif.

Informasi yang dapat ditangkap melalui pada Gambar 5 ada-lah banyak wisatawan merasa senang dan puas ketika me-ngunjungi Candi Borobudur. Adanya kata "*worth*", "*visit*", "*absolutely*", dan "*amaze*" dalam *word cloud* tersebut memi-liki arti yakni Candi Borobudur merupakan destinasi wisata yang layak dikunjungi karena keindahannya. Selain itu wisa-tawan juga berpendapat bahwa Candi Borobudur terawat de-ngan baik, yang ditandai dengan kata "*well*" dan "*clean*". Da-lam *word cloud* bersentimen positif, terlihat adanya kata "*crowd*" yang sebenarnya memiliki makna negatif. Hal terse-but dapat disebabkan kata "*crowd*" muncul beriringan dalam ulasan yang bersentimen positif.

Pada Gambar 5 dan Gambar 6, terdapat dua kata yang memi-liki kesamaan ukuran, yaitu "*visit*" dan "*sunrise*". Artinya, kata "*visit*" dan "*sunrise*" memiliki frekuensi kemunculan yang sama besarnya di kedua kelas sentimen. Kedua kata ter-sebut juga menjadi indikasi bahwa banyak wisatawan yang mengunjungi Candi Borobudur dan tertarik dengan *sunrise tour*. Wisatawan hanya dapat mengakses *sunrise tour* melalui Hotel Manohara.

Meskipun *sunrise tour* merupakan program yang da-pat menarik minat wisatawan, namun tidak sedikit keluhan yang muncul akibat mahalnya harga tiket *sunrise tour*. Selain itu terdapat juga keluhan mengenai ketimpangan harga tiket masuk reguler antara wisatawan domestik dan asing, sebab wisatawan asing perlu membayar dengan harga jauh lebih tinggi. Keluhan tersebut dapat diidentifikasi karena adanya kata "*expensive*", "*overprice*", "*extremely*", "*entrance*", serta "*price*" pada *word cloud* sentimen negatif. Informasi menge-nai hal yang patut menjadi perhatian bagi pihak pengelola wi-sata dapat diperoleh dengan cara memvisualisasikan sepuluh kata dengan frekuensi kemunculan tertinggi pada data ulasan bersentimen negatif dalam bentuk *pareto chart*. Gambar 7 merupakan hasil *pareto chart* tersebut.



Gambar 7. *Pareto Chart Unigram* Sentimen Negatif.

Tanpa melibatkan kata "*visit*" dan "*sunrise*" dalam *pareto chart* tersebut, dapat diperoleh informasi yakni permasalahan keramaian dan harga tiket masuk merupakan hal yang patut mendapatkan perhatian lebih. Hal ini diidentifikasi dari ada-nya kata "*crowd*", "*ticket*", "*entrance*", "*pay*", dan "*price*" yang muncul dalam 80% kumulatif persentase.

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Kesimpulan yang diperoleh dari penelitian ini adalah klasifikasi sentimen wisatawan Candi Borobudur dengan me-tode *K-Nearest Neighbor* dan teknik *unigram* memiliki kiner-ja yang tergolong cukup atau *fair*. Namun ketika digunakan metode *Support Vector Machine kernel Radial Basis Func-tion* dan teknik *unigram*, dihasilkan kinerja yang lebih baik dibandingkan *K-Nearest Neighbor*. Pada penelitian ini, jenis

N-gram terbaik adalah *unigram*, sebab bertambahnya ukuran *n* pada *N-gram* justru menurunkan nilai *Area Under the Curve*. Dari dua metode klasifikasi yang dianalisis, persoalan klasifikasi sentimen wisatawan Candi Borobudur akan lebih baik jika diselesaikan dengan metode *Support Vector Machine kernel Radial Basis Function* dan teknik *unigram*. Hal tersebut dikarenakan metode ini memiliki kinerja klasifikasi yang tergolong sangat baik.

B. Saran

Untuk penelitian selanjutnya, diperlukan ketelitian pa-da tahap *text preprocessing* sebab dapat mempengaruhi hasil klasifikasi. Apabila dalam analisis sentimen dilakukan pelabelan dengan kamus *lexicon* maupun *N-gram*, maka unsur yang perlu ditinjau ulang adalah kata-kata negasi, seperti "*not*", "*no*", "*never*", "*very*", dll. Selain itu pembentukan *feature* dapat dilakukan dengan menerapkan kombinasi *unigram-bigram-trigram*.

Adapun saran yang dapat diberikan bagi pengelola wi-sata Candi Borobudur adalah sebaiknya dilakukan peninjauan ulang mengenai biaya tiket masuk reguler untuk wisatawan asing. Kemudian hal yang perlu dipertahankan ialah tetap diadakan program *sunrise tour*.

DAFTAR PUSTAKA

[1] S. Mukherjee, *Sentiment Analysis of Reviews*, 2nd ed. 2018.

[2] TripAdvisor, "TripAdvisor," 2017. [Online]. Available: https://tripadvisor.mediaroom.com/us-about-us.

[3] Ratman, D. R. (2016). *Pembangunan Destinasi Pariwisata Prioritas 2016-2019.* Jakarta: Kementerian Pariwisata.

[4] Zhang, S., Li, X., Zong, M., Zhu, X., & Wang, R. Efficient kNN Classification with Different Numbers of Nearest Neighbors. *IEEE Transactions on Neural Networks and Learning Systems, 29*(5), 1774-1785. 2018.

[5] Feldman, R., & Sanger, J. *The Text Mining Handbook "Advanced Approaches in Analyzing Unstructured Data"*. Cambridge: University Press. 2007.

[6] Koehn, P. *Statistical Machine Translation.* Cambridge: Cambridge University. 2009.

[7] Marafino, B. J., Davies, J. M., Bardach, N. S., Dean, M. L., & Dudley, R. A. N-gram Support Vector Machines for Scalable Procedure and Diagnosis Classification, with Applications to Clinical Free Text Da-ta from the Intensive Care Unit. *Journal of the American Medical Infor-matics Association, 2-1*(5), 871-875. 2014.

[8] Hung, J., & Zhang, K. Examining Mobile Learning Trends 2003-2008: A Categorical Meta-Trend Analysis Using Text Mining Techni-ques. *Journal of Computing in Higher Education*, 1-17. 2012

[9] Berry, M. W., & Kogan, J. *Text Mining: Applications and Theory 1st Edition.* (Wiley, Penyunt.) United Kingdom. 2010

[10] Liu, B. Sentiment analysis and opinion mining. *Synthesis Lec-tures on Human Language Technologies, 5*(1), 1-167. 2012

[11] Mohammad, S. M., Kiritchenko, S., & Zhu, X. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. *Procee-dings of the Seventh International Workshop on Semantic Evaluation Exercises (SemEval-2013).* Atlanta, USA. 2013

[12] Weiss, S. M. *Text Mining: Predictive Methods for Analyzing.* New York: Springer. 2010

[13] Toman, M., Tesar, R., & Jezek, K. Influence of Word Normalization on Text Classification. *In Proceedings of the 1st International Conference on Multidisciplinary Information Sciences & Technolo-gies, 2*, 354-358. 2006

[14] Liu, B. "Sentiment Analysis and Subjectivity". Dalam Taylor, & Francis, *Handbook of Natural Language Processing, 2nd Edition.* USA: Boca Raton. 2010.

[15] El-Khair, I. A. Term Weighting. *Encyclopedia of Database Systems, 1*, 3037-3040. 2009.

[16] Garreta, R., & Moncecchi, G. *Learning Scikit-Learn: Machine Learning in Python.* Berlin Heidelberg: Packt. 2013.

[17] Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, K. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial In-telligence esearch, 16*, 321-357. 2002.

[18] Sain, H., & Purnami, S. W. Combine Sampling Support Vector Machine fo Imbalanced Data Classification. *Procedia Computer Sci-ence, 72*, 59-66. 2015.

[19] Ma, et al. Integrating Growth and Environmental Parameters to Discriminate Powdery Mildew and Aphid of Winter Wheat Using Bi-Temporal Landsat-8 Imagery. *Remote Sensing, 11*(846), 1-23. 2019.

[20] Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H., & Santos, J. Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches. *IEEE Computational Intelligence Magazine, 13*(4), 59-76. 2018.

[21] Deng, N., Tian, Y., & Zhang, C. *Support Vector Machines: Optimization Based Theory, Algorithms, and Extensions*. Chapman and Hall/CRC. 2012.

[22] Webb, A. R., & Copsey, K. D. *Statistical Pattern Recognition.* Wiley. 2011.

[23] Fletcher, R. *Practical Methods of Optimization.* John Wiley & Sons. 1988.

[24] Vapnik, V. N., & Chervonenkis, A. J. *Theory of Pattern Recognition.* 1974.

[25] Hardle, W. K., Prastyo, D. D., & Hafner, C. M. Support Vector Machines with Evolutionary Model Selection for Default Prediction. Dalam *The Oxford Handbook of Applied Nonparametric and Semipa-rametric Econometrics and Statistics* (hal. 346-373). Oxford University Press. 2014.

[26] Kecman, V. Support Vector Machines – An Introduction. *Stud-Fuzz, 177*, 1-47. 2005.

[27] Zheng, W., Wang, H., Ma, L., & Wang, R. An Improved k-Nearest Neighbor Classifcation Algorithm Using Shared Nearest Neighbor Similarity. *Metallurgical & Mining Industry*(10), P133-137. 2015.

[28] Sun, Y., Wong, A. K., & Kamel, M. S. Classification of Imbalanced Data: A Review. *International Journal of Pattern Recognition and Artificial Intelligence, 23*(4), 687-719. 2009.

[29] Han, J., Kamber, M., & Pei, J. *Data Mining: Concepts and Techniques.* Morgan Kaufmann. 2011.

[30] Ma, C. M., Yang, W. S., & Cheng, B. W. How the Parameters of K-nearest Neighbor Algorithm Impact on the Best Classification Ac-curacy: In Case of Parkinson Dataset. *Journal of Applied Sciences, 14*(2), 171-176. 2014.

[31] Bekkar, M., Djemaa, H. K., & Alitouche, T. A. Evaluation Measure for Models Assessment over Imbalanced Data Sets. *Journal of Information Engineering and Aplications, 3*(10), 27-38. 2013.

[32] Sokolova, M., & Lapalme, G. A Systematic Analysis of Performance Measures for Classification Tasks. *Information Processing and Management, 45*(4), 427-437. 2009.

[33] Castella, Q., & Sutton, C. Word Storm: Multiples of Word Clouds for Visual Comparison of Documents. *Computer Research Repository (CoRR)*. 2013.