ANALISIS DATA ULASAN PENGUNJUNG MENGGUNAKAN *LEXICON BASED, SUPPORT VECTOR MACHINE, RANDOM FOREST* DALAM MENENTUKAN SKALA PRIORITAS PEMBANGUNAN OBJEK WISATA LABUAN BAJO

Tesis

untuk memenuhi sebagian persyaratan

mencapai derajat Sarjana S-2 Program Studi

Magister Sistem Informasi

Arnoldus Janssen Dahur

30000322410011

SEKOLAH PASCASARJANA

UNIVERSITAS DIPONEGORO

HALAMAN PENGESAHAN

TESIS

ANALISIS DATA ULASAN PENGUNJUNG MENGGUNAKAN LEXICON BASED, SUPPORT VECTOR MACHINE, RANDOM FOREST DALAM MENENTUKAN SKALA PRIORITAS PEMBANGUNAN OBJEK WISATA LABUAN BAJO

Oleh:
Arnoldus Janssen Dahur
30000322410011

Telah diujikan dan dinyatakan lulus ujian tesis pada tanggal 23 Januari 2024 oleh tim penguji Program Studi Magister Sistem Informasi Sekolah Pascasarjana Universitas Diponegoro.

Semarang, 23 Januari 2024
Mengetahui,

Penguji I

Dr. Budi Warsito, S.Si., M.Si.
NIP 197508241999031003

Penguji II

Dr. Drs. Catur Edi Widodo, M.T.
NIP 196405181992031002

Pembimbing I

Dr. Eng. Wahyul Amien Syafei, S.T.,M.T.
NIP 197112181995121001

Pembimbing II

Ir. Toni Prahasto, MAsc., Ph.D.
NIP 196208091988031001

Mengetahui :
Dekan Sekolah Pascasarjana
Universitas Diponegoro

Dr. R.B. Sularto, S.H., M.Hum.
NIP 196701011991031005

Ketua Program Studi
Magister Sistem Informasi

Drs. Bayu Surarso, M.Sc., Ph.D.
NIP 19631105198803100l

PERNYATAAN PERSETUJUAN
PUBLIKASI TESIS UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Diponegoro, saya yang bertanda tangan di bawah ini :

Nama	:	Arnoldus Janssen Dahur
NIM	:	30000322410011
Program Studi	:	Magister Sistem Informasi
Program	:	Sekolah Pascasarjana
Jenis Karya	:	Tesis

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Diponegoro Hak Bebas Royalti Noneksklusif atas karya ilmiah saya yang berjudul :

ANALISIS DATA ULASAN PENGUNJUNG MENGGUNAKAN *LEXICON BASED, SUPPORT VECTOR MACHINE, RANDOM FOREST* DALAM MENENTUKAN SKALA PRIORITAS PEMBANGUNAN OBJEK WISATA LABUAN BAJO

beserta perangkat yang ada. Dengan Hak bebas Royalti Noneksklusif ini Program Studi Magister Sistem Informasi Sekolah Pascasarjana Universitas Diponegoro berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*) merawat, dan mempublikasikan tesis saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik hak cipta.

Dibuat di : Semarang
Pada tanggal : 23 Januari 2024
Yang menyatakan

Arnoldus Janssen Dahur
NIM 30000322410011

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam tesis ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar akademik di suatu perguruan tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Semarang, 23 Januari 2023

Arnoldus Janssen Dahur

KATA PENGANTAR

Segala Puji dan syukur penulis panjatkan kehadirat Tuhan Yang Maha Esa yang telah melimpahkan segala rahmatNya sehingga penulis dapat menyelesaikan tesis dengan judul "Analisis Data Ulasan Pengunjung Menggunakan *Lexicon Based, Support Vector Machine, Random Forest* dalam Menentukan Skala Prioritas Pembangunan Objek Wisata Labuan Bajo" guna memenuhi sebagian persyaratan untuk memperoleh gelar magister pada program studi Magister Sistem Informasi, Sekolah Pascasarjana Universitas Diponegoro.

Penulis menyadari kelemahan serta keterbatasan yang ada sehingga dalam menyelesaikan tesis ini memperoleh bantuan dari berbagai pihak, dalam kesempatan ini penulis menyampaikan ucapan terimakasih kepada :

1. Bapak Dr. R.B. Sularto, S.H., M.Hum. selaku dekan Sekolah Pascasarjana Universitas Diponegoro.
2. Bapak Drs. Bayu Surarso, M.Sc., Ph.D. selaku Ketua Program Studi Magister Sistem Informasi Universitas Diponegoro.
3. Bapak Dr. Eng.Wahyul Amien Syafei, S.T.,MT. Selaku Dosen Pembimbing I, yang telah memberikan bimbingan, arahan, serta kritik kepada penulis hingga bisa selesai dalam penyusunan tesis.
4. Bapak Ir. Toni Prahasto, MAsc., Ph.D. selaku dosen Pembimbing II, yang sudah membimbing, memberikan arahan, dan masukan yang baik kepada penulis sehingga dapat penulis menyelesaikan laporan tesis ini.
5. Bapak Dr. Budi Warsito, S.Si., M.Si. selaku Dosen Ketua Penguji yang telah memberikan banyak saran dan masukan yang baik kepada penulis.
6. Bapak Dr. Drs. Catur Edi Widodo, M.T. sebagai Dosen Penguji yang telah memberikan banyak saran dan masukan dalam penyusunan tesis.
7. Semua pihak yang tidak dapat disebutkan satu persatu, yang telah membantu sampai dengan terselesaikannya tesis ini.

Penulis menyadari bahwa tesis ini masih banyak kekurangan baik isi maupun susunannya. Semoga tesis ini dapat bermanfaat tidak hanya bagi penulis juga bagi para pembaca.

Semarang, 30 Januari 2024

Arnoldus Janssen Dahur

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

Daftar Arti Lambang dan Singkatan

Daftar Arti Lambang

Lambang	Arti Lambang
W^T	Parameter Bobot
x	Vector Input
b	bias
$\|\|w\|\|$	Norma Euclidean dari vektor bobot w
D	Jarak support vector dengan hyperplane
$x1, x2$ $...xn$	Data miror acak
$tf_{t,d}$	Jumlah yang muncul di *term* t pada dokumen d
N	Jumlah yang terdapat di dokumen teks
df_t	Jumlah dokumen yang terdapat term t
$W_{tft,d}$	Nilai term frequency
df_t	Jumlah dokumen yang terdapat term
$Recall$	Rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif
$Precision$	Rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif
$log_{10}\frac{N}{df_t}$	Log dari jumlah keseluruhan dokumen dibagi dengan jumlah kata yang sama pada semua dokumen

Daftar Singkatan

Singkatan	Kepanjangan Singkatan
TP	*True Positive*
TN	*True Negative*
FP	*False Positive*
FN	*False Negative*
HTML	*HyperText Markup Language*
HTML	*Extensible HyperText Markup Language*
CSV	*Comma-separated- values*
SVM	*Support Vector Machine*
RF	*Random Forest*
ROC	*Receiver Operating Characteristic*
AUC	*Area Under the Curve*
CRISP-DM	*Cross-Industry Standard Process for Data Mining*
TF-IDF	*Term Frequency-Inverse Document Frequency*

ANALISIS DATA ULASAN PENGUNJUNG *MENGGUNAKAN LEXICON BASED, SUPPORT VECTOR MACHINE, RANDOM FOREST* DALAM MENENTUKAN SKALA PRIORITAS PEMBANGUNAN OBJEK WISATA LABUAN BAJO

ABSTRAK

Objek wisata Labuan Bajo merupakan salah satu destinasi wisata super prioritas di Indonesia. Pentingnya mendapatkan dan menganalisis ulasan pengunjung wisata untuk mengetahui preferensi berupa pandangan pengunjung terhadap fasilitas dan pelayanan yang ada saat ini. Oleh karena itu penelitian ini dilakukan untuk untuk mendapatkan dan menganalisis data ulasan pengunjung yang didapat dari website TripAdvisor dan Google Maps. Adapun metode yang digunakan dalam analisis ulasan pengunjung ini yaitu Lexicon Based untuk melakukan pelabelan, metode Support Vector Machine (SVM) dan Random Forest untuk klasifikasi. Hasil pelabelan menggunakan metode Lexicon Based didapat sentiment positif sejumlah 4187 ulasan, sentiment negatif sejumlah 1796 ulasan dan sentiment netral sejumlah 1774 ulasan. Klasifikasi dilakukan dengan menggunakan teknik *undersampling* dan tanpa menggunakan teknik *undersampling* karena ketidak seimbangan data. Hasil menggunakan teknik *undersampling* dengan SVM yaitu *accuracy* 0.89 *precisi* 0.95 recall 0.85 dan *f1-measure* 0.90 serta nilai ROC AUC menggunakan teknik *undersampling* yaitu 0.94 dan tanpa menggunakan teknik *undersampling accuracy* 0.79 presisi 0.80 recall 0.94 dan *f1-measure* 0.86 serta nilai ROC AUC yaitu 0.83. Hasil menggunakan teknik *undersampling* dengan Random Forest yaitu *accuracy* 0.87 *precisi* 0.91 recall 0.86 dan *f1-measure* 0.88 serta nilai ROC AUC menggunakan teknik *undersampling* yaitu 0.93 dan tanpa menggunakan SMOTE *accuracy* 0.77 presisi 0.78 recall 0.94 dan *f1-measure* 0.85 serta nilai ROC AUC yaitu 0.81. Penentuan skala prioritas dilakukan dengan mendapatkan 10 kata teratas dan jumlah sentiment dari masing-masing kata yang berkaitan dengan pembangunan didapat kata-kata sentimen positif yang sering muncul yaitu ’indah’, ’alami’, ’eksotik’,’pandang’, ’bersih’,’ purba’, ’takjub’, ’sejarah’. Pelestarian aset alami dan aset sejarah tentunya harus dijaga dan terus dipertahankan. Sebaliknya kata-kata negatif yang sering muncul yaitu ’mahal’, ’biaya’, ’pandu’, ’jalan’, ’sampah’, ’panas’. Berdasarkan kata tersebut pembangunan transportasi dan infrastruktur tentunya sangat diperlukan dalam peningkatan daya tarik wisata Labuan Bajo.

Kata kunci: Analisis Ulasan, Labuan Bajo, Lexicon Based, Support Vector Machine, Random Forest

ANALYSIS OF VISITOR REVIEW DATA USING *LEXICON BASED, SUPPORT VECTOR MACHINE, RANDOM FOREST* IN DETERMINING THE PRIORITY SCALE OF BUILDING LABUAN BAJO TOURISM OBJECTS

ABSTRACT

Labuan Bajo tourist destination is one of the super priority tourist destinations in Indonesia. The importance of obtaining and analyzing tourists' reviews is to understand their preferences and views on the existing facilities and services. Therefore, this research is conducted to obtain and analyze visitor review data obtained from TripAdvisor and Google Maps. The methods used in analyzing these visitor reviews are Lexicon-Based for labeling, Support Vector Machine (SVM), and Random Forest for classification. The labeling results using the Lexicon-Based method showed 4187 positive reviews, 1796 negative reviews, and 1774 neutral reviews. The classification was performed using undersampling technique and without using undersampling techniquedue to data imbalance. Results using undersampling technique with SVM showed an accuracy of 0.89, precision of 0.95, recall of 0.85, and f1-measure of 0.90, with an ROC AUC value of 0.94. Without using undersampling technique, the accuracy was 0.79, precision was 0.80, recall was 0.94, and f1-measure was 0.86, with an ROC AUC value of 0.83. Results using undersampling techniquewith Random Forest showed an accuracy of 0.87, precision of 0.91, recall of 0.86, and f1-measure of 0.88, with an ROC AUC value of 0.93. Without using undersampling technique, the accuracy was 0.77, precision was 0.78, recall was 0.94, and f1-measure was 0.85, with an ROC AUC value of 0.81.The determination of priority scale was done by obtaining the top 10 words and the number of sentiments related to development. The frequently occurring positive sentiment words were 'beautiful,' 'natural,' 'exotic,' 'scenic,' 'clean,' 'ancient,' 'amazed,' and 'historical.' The preservation of natural and historical assets must be maintained and continuously preserved.On theother hand, the frequently occurring negative words were 'expensive,' 'cost,' 'guide,''road,' 'garbage,' and 'hot.' Based on these words, the development of transportationand infrastructure is undoubtedly needed to enhance the attractiveness of Labuan Bajo as a tourist destination.

Keywords: Review Analysis, Labuan Bajo, Lexicon Based, Support Vector Machine, Random Forest.