ANALISIS ULASAN HOTEL DI SITUS TRIPADVISOR MENGGUNAKAN METODE *TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY* DAN *K-NEAREST NEIGHBOR*

Tesis untuk Tesis S-2

Program Studi Magister Sistem Informasi

Khairul Huda
30000318410014

PROGRAM PASCASARJANA

UNIVERSITAS DIPONEGORO

SEMARANG

2022

HALAMAN PENGESAHAN

TESIS

ANALISIS ULASAN HOTEL DI SITUS TRIPADVISOR MENGGUNAKAN METODE *TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY* (TF-IDF) DAN *K-NEAREST NEIGHBOR* (KNN)

Oleh:
Khairul Huda
30000318410014

Telah diujikan dan dinyatakan lulus ujian tesis pada tanggal 15 Agustus 2022 oleh tim penguji Program Studi Magister Sistem Informasi Sekolah Pascasarjana Universitas Diponegoro.

Semarang, 15 Agustus 2022
Mengetahui,

Penguji I	**Penguji II**
Prof. Dr. Kusworo Adi, S.Si., M.T.	Drs. Bayu Surarso, M.Sc., Ph.D
NIP: 197203171998021001	NIP. 196311051988031001
Pembimbing I	**Pembimbing II**
Dr. Catur Edi Widodo, MT.	Vincencius Gunawan S.K., M.Si., Ph.D
NIP: 196405181992031002	NIP: 197105221997021001

Mengetahui : **Dekan Sekolah Pascasarjana** **Universitas Diponegoro**	**Ketua Program Studi** **Magister Sistem Informasi**
Dr. R.B. Sularto, S.H., M.Hum	Drs. Bayu Surarso, M.Sc., Ph.D
NIP. 196701011991031005	NIP. 196311051988031001

PERNYATAAN PERSETUJUAN
PUBLIKASI TESIS UNTUK KEPENTINGAN AKADEMIS

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

Nama	:	Khairul Huda
NIM	:	30000318410014
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 ULASAN HOTEL DI SITUS TRIPADVISOR MENGGUNAKAN METODE *TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY* DAN *K-NEAREST NEIGHBOR*

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: 12 Juli 2022
Yang menyatakan

Khairul Huda
NIM. 30000318410014

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, 12 Juli 2022

Khairul Huda

KATA PENGANTAR

Puji syukur ke hadirat Allah SWT, atas limpahan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan penelitian tesis dengan judul "Analisis Ulasan Hotel Di Situs TripAdvisor Menggunakan Metode *Term Frequency-Inverse Document Frequency* (TF-IDF) Dan *K-Nearest Neighbor* (KNN)". Penulis menyadari tidak dapat menyelesaikan Tesis ini tanpa bantuan dan dorongan dari berbagai pihak. Untuk itu pada kesempatan ini penulis ingin mengucapkan terima kasih kepada:

1. Bapak Dr. Catur Edi Widodo, MT selaku pembimbing pertama penulis yang telah memberikan wawasan, masukan, arahan dan nasehat dalam penulisan tesis.
2. **Bapak** Vincencius Gunawan S.K., M.Si., Ph.D selaku pembimbing kedua yang juga telah memberikan bimbingan, koreksi dan dukungan dalam penulisan tesis.
3. Bapak Drs. Bayu Surarso, M.Sc., Ph.D **Selaku Ketua Program Studi, Magister Sistem Informasi, Sekolah Pascasarjana, Universitas Diponegoro, Semarang.**
4. Bapak **Dr. R.B. Sularto, S.H., M.Hum, Selaku Dekan Sekolah Pascasarjana Universitas Diponegoro Semarang.**
5. Keluargaku tercinta, Bapak Drs. Pardi dan Ibu Sri Wahyuti, serta istri dan kakak-kakak saya.
6. Seluruh Dosen Magister Sistem Informasi, Universitas Diponegoro yang tidak dapat penulis sebutkan satu persatu, terimakasih atas ilmu yang diberikan selama perkuliahan.
7. Rekan-rekan Magister Sistem Informasi 2018 dan semua staf karyawan Pascasarjana, Universitas Diponegoro, terimakasih atas kerja sama yang telah terjalin selama ini, sehingga penulis menyelesaikan administrasi hingga menyelesaikan tesis ini.

Penulis menyadari bahwa tesis ini masih jauh dari sempurna, untuk itu saran dan kritik yang membangun sangat penulis harapkan. Semoga tesis ini dapat bermanfaat bagi pihak-pihak yang membutuhkan.

Semarang, Juni 2022
Penulis

DAFTAR ISI

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR TABEL

Halaman

DAFTAR LAMPIRAN

Halaman

DAFTAR ARTI LAMBANG DAN SINGKATAN

DAFTAR ARTI LAMBANG

Lambang	Arti Lambang
D_{xy}	Jarak antara objek x dan y
Cos_{ij}	Nilai kemiripan data uji dengan data latih
d_{ik}	Panjang vektor data uji
d_{jk}	Panjang vektor data latih
x	Data uji yang akan diklasifikasi
y	Data latih
n	Jumlah data latih
i	Data ke-i
Tf	Banyaknya kata dalam dokumen
$Tf_{t,d}$	Banyaknya kata *t* pada dokumen d
d	Jumlah dokumen
df	Jumlah dokumen yang mengandung suatu *term*
$W_{t,d}$	Nilai bobot akhir kata
$tf_{t,d}$	Banyaknya kata dalam dokumen
idf_t	Pembobotan keseluruhan

DAFTAR SINGKATAN

Singkatan	Kepanjangan Singkatan
SVM	*Support Vector Machine*
ANN	*Artificial Neural Network*
KNN	*K-Nearest Neighbor*
CNN	*Convolutional Neural Network*
SQL	*Structured Query Language*
TF-IDF	*Term Frequency-Inverse Document Frequency*
KDD	*Knowledge Discovery in Database*
NLTK	*Natural Language Toolkit*
API	*Application Programming Interface*
NLP	*Natural Languange Processing*
UML	*Unified Modelling Language*
SDLC	*Software Development Language Cycle*

ANALISIS ULASAN HOTEL DI SITUS TRIPADVISOR MENGGUNAKAN METODE *TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY* DAN *K-NEAREST NEIGHBOR*

ABSTRAK

Penelitian ini dilatarbelakangi oleh evaluasi produk dan jasa menggunakan metode konvensional melalui wawancara, survei dan kuisioner yang berakibat pada hasil analisis menjadi tidak akurat dan tidak konsisten. Penelitian ini bertujuan untuk menerapkan algoritma *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *K-Nearest Neighbor* serta mengevalusi hasil dari sistem yang dibangun dengan tingkat akurasi yang paling optimal. Salah satu upaya untuk menanggulangi permasalahan tersebut yaitu membangun sistem untuk analisis ulasan pelanggan hotel di situs TripAdvisor yang bernilai positif, negatif dan netral menggunakan teknik *text mining.* Algoritma klasifikasi yang digunakan dalam penelitian adalah *K-Nearest Neighbor* karena memiliki kelebihan dalam hal komputasi berkinerja tinggi, tahan terhadap berbagai karakteristik data yang besar, dan memiliki kompleksitas algoritma yang relatif kecil. Hasil penelitian menunjukkan bahwa sistem dapat melakukan klasifikasi terhadap ulasan hotel di situs TripAdvisor yang bernilai positif, negatif dan netral dengan nilai performa paling baik pada K=31 dan memiliki tingkat akurasi mencapai 76% untuk data *training*, dan menghasilkan peningkatan akurasi mencapai 84% dengan menerapkan metode *random over-sampling* untuk *rebalance* data.

Kata kunci : *text mining, term frequency-inverse document frequency, k-nearest neighbor, random over-sampling*

ANALYSIS OF HOTEL REVIEWS ON THE TRIPADVISOR SITE USING THE TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY AND K-NEAREST NEIGHBOR METHODS

ABSTRACT

This research is based on analysis using conventional methods through interviews, surveys and questionnaires which resulted in the analysis being inaccurate and inconsistent. This study aims to apply the Term Frequency-Inverse Document Frequency (TF-IDF) and K-Nearest Neighbor algorithms and evaluate the results of the system built with the most optimal level of accuracy. To solve these problems is to build a system for analyzing hotel customer reviews on the TripAdvisor site which have positive, negative and neutral values using text mining techniques. The classification algorithm used in the research is K-Nearest Neighbor because it has advantages in terms of high-performance computing, is resistant to various characteristics of large data, and has a relatively small algorithm complexity. The results show that the system can classify hotel reviews on the TripAdvisor site which are positive, negative and neutral with the best performance value at K = 31 and has an accuracy of 76% for training data and increase in accuracy of up to 84% by applying the random over-sampling method for data rebalance.

Keywords: text mining, term frequency-inverse document frequency, k-nearest neighbor, random over-sampling