PENERAPAN METODE SUPPORT VECTOR MACHINE UNTUK ANALISIS SENTIMEN PADA ULASAN PELANGGAN HOTEL DI TRIPADVISOR

Willyanto Wijaya [1] Dyah Erny Herwindiati [2] Novario Jaya Perdana [3]

[1) 2) 3)] Teknik Informatika, FTI, Universitas Tarumanagara
Jl. Letjen S Parman no 1, Jakarta 11440 Indonesia
[1)]email : willyanto.535180017@stu.untar.ac.id, [2)]email : dyahh@fti.untar.ac.id, [3)]email : novariojp@fti.untar.ac.id

ABSTRACT

Indonesia is an archipelagic country that has very beautiful nature, in addition to the natural beauty of cultural diversity is also one of the factors Indonesia has a tourist attraction. One of the effects of Indonesia's natural beauty and cultural diversity can be seen from the increase in hotel occupancy rates. This hotel analysis system design uses training data and test data from the tripadvisor website. Tripadvisor is a website that focuses on tourism. on tripadvisor there are a lot of services offered ranging from transportation, lodging, travel experiences, and restaurants. One of the useful features of tripadvisor is the review column, this review column can be used to do research. visitor reviews from the tripadvisor comments column can be used as a value. to visualize and see people's emotions how the services provided by the hotel to visitors. The research phase starts from scrapping data from the triapadvisor review column, preprocessing data, word weighting, SVM, and evaluation with a confusion matrix. The data taken from the review column is done by web scraping technique. This study uses data from 3000 reviews from 15 hotels. The results of the classification will then be evaluated with a confusion matrix. The highest accuracy result will be used as a model for classification. the classification results will be displayed in the form of detailed tables and diagrams that describe the percentage of sentiment classification results.

Keywords: *Confusion Matrix, Preprocessing, Support Vector Machine, Tripadvisor.*

1. Pendahuluan

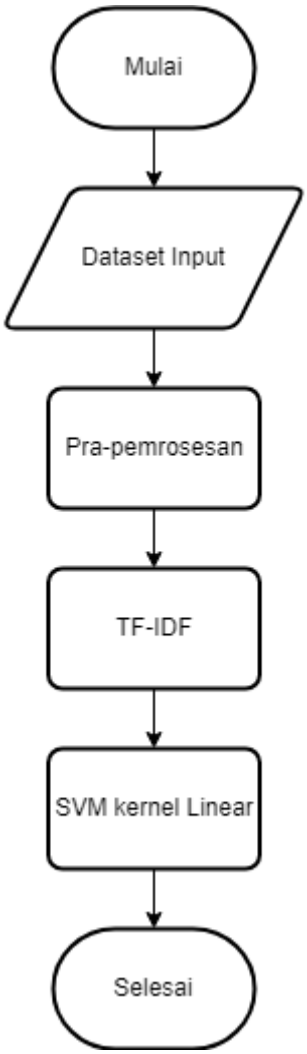
Indonesia merupakan negara kepulauan yang memiliki alam yang sangat indah, selain keindahan alam keragaman budaya juga menjadi salah satu faktor Indonesia memiliki daya tarik wisatawan. Tahun 2019 Indonesia menduduki peringkat ke 29 untuk tingkat kedangan turis internasional [1]. Hal inilah yang membuat industri pariwisata menjadi faktor pendukung pertumbuhan bisnis di Indonesia terutama sektor akomodasi penginapan seperti villa, hotel, resort, dan sejenisnya. Tingkat okupansi hotel di Indonesia pada tahun 2019 adalah sebesar 54,81% [2], dengan perbandigan wisatawan lokal sebanyak 86,75% [3] dan wisatawan asing sebanyak 13,24% [4]. Menurut organisasi wisata dunia atau WTO wisatawan biasanya menginap minimal 24 jam atau maksimal 6 bulan [5], hal inillah yang membuat wisatawan membutuhkan hotel atau penginapan untuk ditempati dalam beberapa waktu. Saat ini dalam melakukan pencarian dan pemesanan hotel wisatawan biasanya menggunakan aplikasi secara daring lewat gawai yang mereka miliki karena menghemat waktu dan lebih mudah membandingkan harga antar hotel. Penelitian yang dilakukan oleh Yohann Indra tahun 2020 menunjukan bahwa terdapat 5 faktor penting yang menjadi preferensi wisatawan dalam menentukan hotel yang akan mereka tempati yaitu: kenyamanan, value for money, reputasi dan pelayanan, fasilitas, dan makanan [6]. Hal ini membuktikan bahwa ulasan yang diberikan pengunjung sebelumnya terhadap hotel memiliki dampak terhadap pemilihan hotel bagi wisatawan. Perancangan ini akan menganalisis kolom ulasan dari web Tripadvisor sebagai sarana untuk melihat pandangan dan pendapat pengunjung terhadap pihak hotel.

Penelitian yang sudah ada sebelumnya tentang sentimen analisis wacana pemindahan ibu kota negara menggunakan algoritma support vector machine (SVM) didapatkan akurasi sebesar 96,68%, *precision* sebesar 95,82%, dan *recall* sebesar 94,04% [7]. Kemudian penelitian terhadap stay home didapatkan akurasi untuk naïve bayes sebesar 66,81%, support vector machine sebesar 80,05% dan k-nearest neighbor sebesar 51,45% [8]. Oleh karena tingkat akurasi algoritma support vector machine (SVM) yang dihasilkan lebih baik dibandingkan algoritma lainnya maka dipilihlah algoritma support vector machine (SVM). Klasifikasi menggunakan algoritma SVM kernel linear, Berdasarkan jurnal bisnis manajemen dan informatika yang dibuat oleh Faizal Fakhri Irfani tahun 2020 klasifikasi kernel support vector machine terbaik yaitu kernel linear sebesar 89,7% dibandingkan kernel rbf 86,5%, dan polynomial 84,5%. Oleh sebab itu dipilihlah Support Vector Machine kernel

linear untuk pengujian.[9]. Karena mendapatkan akurasi terbesar kernel linear diharapkan dapat menentukan garis *hyperplane* terbaik antara dua kelas yaitu positif dan negatif [10]. Sehingga dapat digunakan untuk melakukan prediksi kelas yang didasari proses training. Hasil dari analisis akan divisualisasikan ke dalam bentuk diagram supaya data atau informasi yang dihasilkan akan lebih mudah untuk dipahami. Hasil klasifikasi akan ditampilkan dalam bentuk pie chart, diagram garis, dan table detil.

2. Metode

Terdapat beberapa tahapan yang dilakukan pada penelitian ini. Alur tahapan tersebut dapat dilihat pada **Gambar 1**.



Gambar 1 Diagram Alir

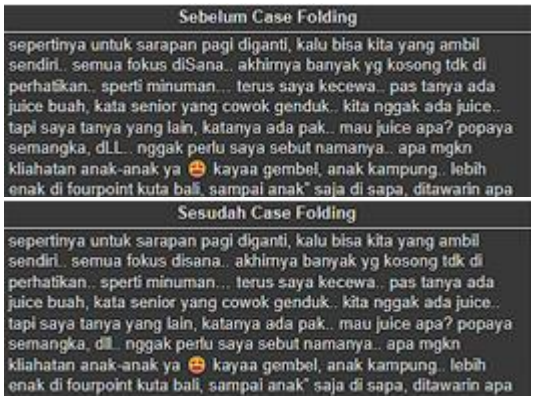
2.1. Data Input

Data input yang digunakan adalah data ulasan yang berasal dari kolom ulasan website Tripadvisor. Total data yang diambil adalah sebanyak 3000 ulasan dari 15 hotel. Pengambilan data dilakukan dengan menggunakan *library beautifulsoup* dengan bahasa pemrograman python.

Setelah mendapatkan data input langkah selanjutnya adalah proses pra-pemrosesan, tahap pra-pemrosesan dilakukan untuk membersihkan data agar dapat diolah lebih lanjut. Ada beberapa tahap pra-pemrosesan yaitu:

1. Case Folding

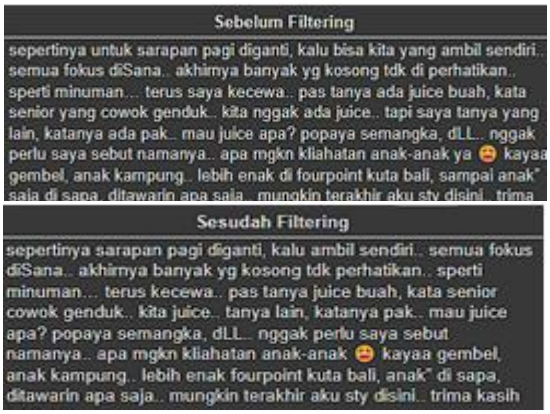
Tahap ini dilakukan untuk menyeragamkan karakter pada data. Pada proses ini semua huruf kapital diubah menjadi huruf kecil. Contohnya dapat dilihat pada **Gambar 2**.



Gambar 2 Penerapan Case Folding

2. Filtering

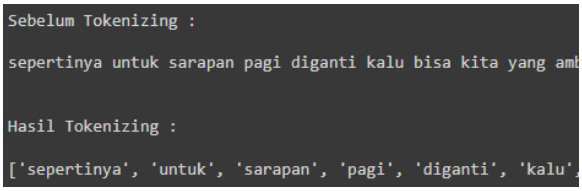
Tahap ini dilakukan untuk menghapus kata yang kurang penting dan tidak memiliki arti. Contohnya di **Gambar 3** dimana kata “di”, “ bisa”, “kita” dihilangkan karena kurang penting.



Gambar 3 Penerapan Filtering

3. Tokenizing

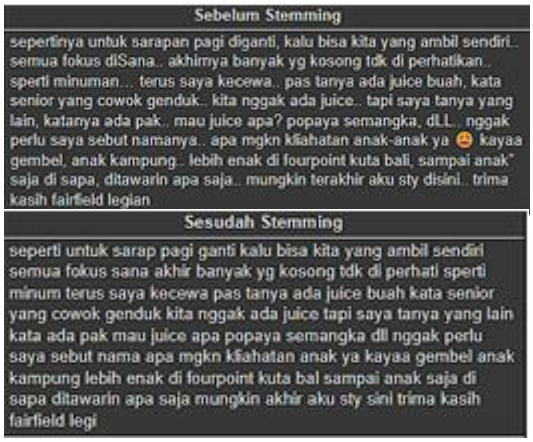
Tahap ini dilakukan untuk membagi kata-kata menjadi sebuah token-token agar setiap kata memiliki nilai sendiri. Contoh penerapannya dapat dilihat pada **Gambar 4**.



Gambar 4 Penerapan Tokenizing

4. Stemming

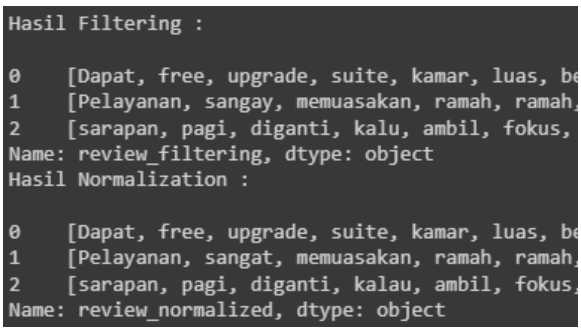
Tahap ini dilakukan untuk menghapus imbuhan yang terdapat pada suatu kata baik awalan, sisipan, maupun gabungan sehingga kata tersebut menjadi kata dasar Kembali. Contohnya dapat dilihat pada **Gambar 5**.



Gambar 5 Penerapan Stemming

5. Normalization

Tahap ini dilakukan untuk menyeragamkan kata yang mempunyai makna namun penulisannya berbeda yang dapat diakibatkan salah ketik dan bahasa gaul. Contoh penerapannya dapat dilihat pada **Gambar 6**.



Gambar 6 Penerapan Normalization

2.2. Pembobotan TF-IDF

Tahapan TF berfungsi adalah untuk menghitung jumlah kemunculan kata pada suatu dokumen. *Term frequency* berfungsi untuk menentukan bobot suatu kata dalam sebuah dokumen berdasarkan banyak munculnya suatu kata didalam suatu dokumen. Metode pembobotan persamaan *Term Frequency(TF)* dan *Inverse Document Frequency (IDF)* dapat dilihat pada **persamaan(1).**

$$W_{dt}=TF_{dt}*IDF_{ft} \quad (1)$$

TF-IDF ini berfungsi untuk menentukan representasi nilai dari kumpulan data training yang akan dibentuk vektor antara dokumen dan kata disatukan berdasarkan kesamaan antara dokumen dan kata.

2.3. Support Vector Machine

Support Vector Machine dikenal sebagai Teknik pembelajaran mesin paling mutakhir setelah pembelajaran mesin sebelumnya yaitu *Neural Network.* Support Vector Machine melakukan pembelajaran dengan menggunakan pasangan data *input* dan data *output* berupa sasaran yang diinginkan. Support Vector Machine secara sederhana dapat dijelaskan sebagai usaha untuk mencari *hyperplane* terbaik yang berfungsi untuk pemisah antar dua kelas input. Perancangan ini menggunakan Support Vector Machine kernel linear sebagai model prediksinya dengan menggunakan fungsi $f(w,b)=Xi*w+b$ untuk menemukan kelas pemisah dengan hasil 1 dan -1 dapat dilihat pada **persamaan(2) dan persamaan (3)**.

$$W*X_i+b \geq +1 \; untuk \; Y_i=+1 \quad (2)$$

$$W*X_i+b \leq -1 \; untuk \; Y_i=-1 \quad (3)$$

Dimana:

Xi = inisialisasi data ke-i

W = nilai support vector yang sebanding dengan hyperplane

b = nilai hasil bias

Yi = data ke- yang nilai ekuivalen dengan persamaan 4

$$Y_i(X_i*W+b)-1+0 \; untuk \; i=1,..,n \quad (4)$$

Dimana:

n = nilai seluruh sampai data ke n

3. Hasil

Pengujian dilakukan menjadi 2 tahap yaitu pengujian metode dan tampilan website. Data latih dan uji yang digunakan didapatkan dari kolom ulasan 15 hotel di 5 provinsi dengan jumlah data perkelas yang dapat dilihat pada **Tabel 1**.

Tabel 1 Jumlah data ulasan per kelas

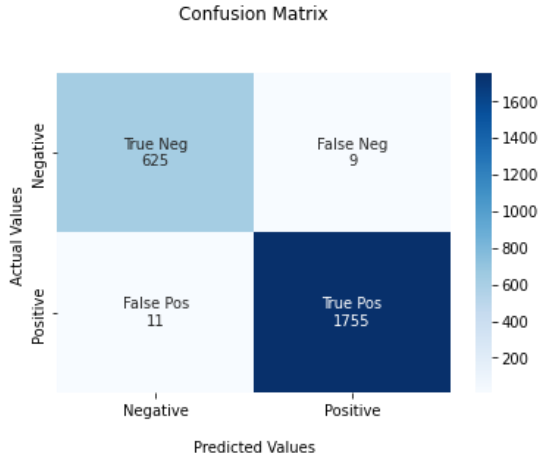
Kelas	Jumlah
Positive	2202
Negative	798
Jumlah Data	**3000**

Untuk pengujian data latih dan data uji, dilakukan pengujian sebanyak 10 kali, dengan melakukan pembagian 80% data latih dan 20% data uji. Pengujian ini dilakukan dengan menggunakan algoritma Support Vector Machine kernel linear yang dapat dilihat detilnya pada **Tabel 2**.

Tabel 2 Pengujian data latih

Pengujian	Akurasi	F1 Score	Precision	Recall
1	0.989	0.989	0.989	0.989
2	0.990	0.990	0.990	0.990
3	0.989	0.989	0.989	0.989
4	0.988	0.988	0.988	0.988
5	0.990	0.990	0.990	0.990
6	0.991	0.991	0.991	0.991
7	0.989	0.989	0.989	0.989
8	0.989	0.989	0.989	0.989
9	0.992	0.992	0.992	0.992
10	0.990	0.990	0.990	0.990

Dari 10 kali percobaan, didapatkan hasil terbaik pada pengujian ke 9 dengan akurasi sebesar 99,2%%, f1 score 99,2%, precision 99,2%, dan recall 99,2% dengan jumlah data latih sebanyak 2400 ulasan yang terprediksi *True Negative* 625 data, *False Negative* 9 data, *True Positive* 1755 data, dan *False Positive* 11 data yang dapat dilihat detilnya pada **Gambar 7**.



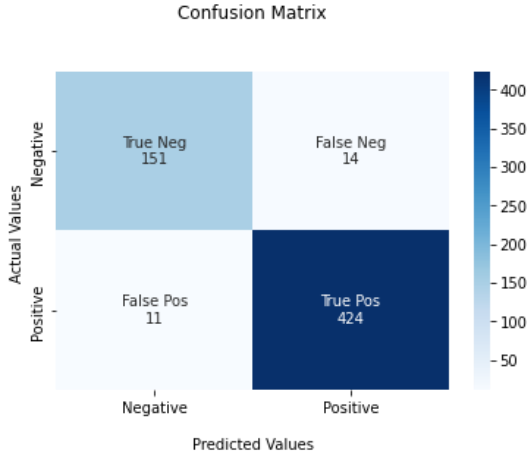
Gambar 7 Confusion Matrix data latih

Tabel 3 Pengujian data uji

Pengujian	Akurasi	F1 Score	Precision	Recall
1	0.947	0.947	0.947	0.947
2	0.947	0.947	0.947	0.947
3	0.948	0.949	0.949	0.948
4	0.958	0.958	0.958	0.958
5	0.947	0.946	0.946	0.947
6	0.943	0.943	0.943	0.943
7	0.953	0.953	0.954	0.953
8	0.937	0.936	0.936	0.937
9	0.942	0.941	0.941	0.942
10	0.938	0.938	0.938	0.938

Dari 10 kali percobaan, didapatkan hasil terbaik pada pengujian ke 4 dimana akurasi sebesar 95,8%, f1 score 95,8%, precision 95,8%, dan recall 95,8%. Hasil pengujian ke 4 inilah yang akan digunakan sebagai model untuk klasifikasi sentimen.

Jumlah data uji yang digunakan adalah sebanyak 600 ulasan atau 20% dari total keseluruhan data yang terprediksi *True Negative* 625 data, *False Negative* 9 data, *True Positive* 1755 data, dan *False Positive* 11 data yang dapat dilihat detilnya pada **Gambar 8**.



Gambar 8 Confusion Matrix data uji

Selanjutnya adalah melakukan pengujian waktu pengambilan ulasan dari web tripadvisor. Pengujian waktu pengambilan dilakukan masing-masing 5 kali percobaan terhadap 5 hotel dengan pengambilan 100, 150, dan 200 ulasan. Hal ini dilakukan untuk mengukur waktu optimal yang diperlukan oleh sistem dalam mengambil ulasan. Waktu pengambilan dapat dilihat pada **Tabel 4, Tabel 5, Tabel 6, Tabel 7, dan Tabel 8**.

Tabel 4 Waktu pengambilan Hotel Grand Mirage (Detik)

Percobaan ke-	100 Ulasan	150 Ulasan	200 Ulasan
1	41.57	69.54	94.00
2	44.42	70.89	92.61
3	43.47	69.90	92.22
4	42.47	69.50	95.44
5	38.16	68.08	94.82
Rata-rata	42.01	69.58	93.82
Waktu per Ulasan	0.42	0.46	0.47

Tabel 5 Waktu pengambilan Hotel Sofitel Bali Nusa Dua (Detik)

Percobaan ke-	100 Ulasan	150 Ulasan	200 Ulasan
1	46.50	68.03	92.51
2	50.20	68.09	93.60
3	46.52	69.16	92.24
4	46.00	69.93	91.45
5	46.48	68.30	91.86
Rata-rata	47.14	68.70	92.33
Waktu per Ulasan	0.47	0.46	0.46

Tabel 6 Waktu pengambilan Hotel Ibis Bandung Pasteur (Detik)

Percobaan ke-	100 Ulasan	150 Ulasan	200 Ulasan
1	41.37	69.81	92.08
2	42.50	68.89	94.61
3	43.82	67.90	92.45
4	42.84	67.60	93.74
5	41.31	67.08	92.89
Rata-rata	42.37	68.26	93.15
Waktu per Ulasan	0.42	0.46	0.47

Tabel 7 Waktu pengambilan Hotel Ibis Styles Malang (Detik)

Percobaan ke-	100 Ulasan	150 Ulasan	200 Ulasan
1	40.91	70.54	96.09
2	41.34	69.45	93.64
3	45.20	68.12	95.34
4	44.06	67.60	94.67
5	41.20	66.98	93.78
Rata-rata	42.54	68.54	94.70
Waktu per Ulasan	0.43	0.46	0.47

Tabel 8 Waktu pengambilan Zest Hotel Bogor (Detik)

Percobaan ke-	100 Ulasan	150 Ulasan	200 Ulasan
1	42.63	68.64	95.41
2	44.80	69.82	97.63
3	44.12	67.41	95.52
4	53.40	66.70	93.94
5	45.20	68.28	94.83
Rata-rata	46.03	68.17	95.47
Waktu per Ulasan	0.46	0.45	0.48

Dari 5 kali percobaan terhadap 5 hotel didapatkan waktu rata-rata untuk 100 ulasan adalah 0,44 per ulasan, untuk 150 ulasan 0,46 dan untuk 200 ulasan 0,47. Dimana hotel Ibis Bandung dan hotel Grand Mirage mempunyai waktu terbaik untuk 100 ulasan yaitu 0,42 detik per ulasan, karena waktu per ulasan untuk pengambilan 100 ulasan dari 5 kali percobaan terhadap 5 hotel paling efisien, maka sistem akan menggunakan sebanyak 100 ulasan untuk proses analisis.

Setelah melakukan tahap pengujian data latih dan data uji tahap selanjutnya adalah melakukan pengujian klasifikasi menggunakan model yang sudah didapatkan dari pengujian data uji ke 4. Model data uji ke 4 dipilih karena mendapatkan nilai terbaik dibanding ke 9 pengujiaan lainnya. Proses pengujian klasifikasi dilakukan terhadap 5 hotel yang berbeda. Hasil dari pengujian klasifikasi dapat dilihat pada **Tabel 9.**

Nama Hotel	Akurasi	F1 Score	Precision	Recall
Pullman Jakarta Central Park	65%	62%	61%	65%
Sofitel Bali Nusa Dua	65%	67%	70%	65%
Hotel Ciputra Jakarta	70%	78%	90%	70%
Padma Hotel Bandung	71%	82%	96%	71%
Sunan Hotel Solo	71%	64%	64%	71%

Dari hasil klasifikasi terhadap 5 hotel didapatkan hasil akurasi terbaik sebesar 71% di 2 hotel yaitu Padma Hotel Bandung dan Sunan Hotel Solo. Namun untuk nilai f1 score, precision, dan recall terbesar didapatkan di Padma Hotel Bandung. Kemudian untuk akurasi terendah didapatkan di hotel Pullman Jakarta Central Park dan Sofitel Bali Nusa Dua yaitu sebesar 65%.

4. Kesimpulan

Dari hasil penelitian ini didapatkan kesimpulan bahwa hasil pengujian data uji terbaik didapatkan pada pengujian ke 4 dengan akurasi sebesar 95,8% yang akan digunakan sebagai model pengklasifikasian. Penerapan algoritma Support Vector Machine kenel linear dengan model pengklasifikasian terbaik mendapatkan akurasi sebesar 71%, 82% f1 score, 96% precision, dan 71% recall. Hasil pengujian ini juga membuktikan bahwa penerapan algoritma Support Vector Machine kernel linear dalam mengklasifikasi teks bekerja dengan cukup baik dan dapat dilihat bahwa sentimen masyarakat terhadap pelayanan yang diberikan pihak hotel cenderung positif.

REFERENSI

[1] UNWTO. "UNWTO World Tourism Barometer and Statistical Annex, December 2020". Jurnal World Tourism Barometer. Vol. 18, Nomor 7. Desember 2020.

[2] Badan Pusat Statistik (BPS). Tingkat Penghunian Kamar Hotel (Persen), 2018-2020. https://www.bps.go.id/indicator/16/282/1/tingkat-penghunian-kamar-hotel.html, 26 Maret 2022.

[3] Badan Pusat Statistik (BPS). Jumlah Tamu Indonesia Pada Hotel Bintang (Ribu Orang), 2018-2020. https://www.bps.go.id/indicator/16/328/1/jumlah-tamu-indonesia-pada-hotel-bintang.html, 26 Maret 2022.

[4] Badan Pusat Statistik (BPS), Jumlah Tamu Asing Pada Hotel Bintang (Ribu Orang), 2018-2020. https://www.bps.go.id/indicator/16/310/1/jumlah-tamu-asing-pada-hotel-bintang.html, 26 Maret 2022.

[5] UNWTO. Glossary Of Tourism Terms. https://www.unwto.org/glossary-tourism-terms, 26 Maret 2022.

[6] Indra, Yohann; Angelina, Tjong; dan Sienny, Thio. "PREFERENSI WISATAWAN SENIOR DALAM MEMILIH HOTEL DI MALANG DAN/ATAU BATU". Jurnal Hospitality dan Manajemen Jasa. Vol.8, Nomor 1, 2020.

[7] Arsi, Primandani; dan Waluyo Retno. "ANALISIS SENTIMEN WACANA PEMINDAHAN IBU KOTA INDONESIA MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE (SVM)". Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK). Vol. 8, Nomor 1, 2021.

[8] Hakim, Ikhwanul; Nugroho, Arifin; Sukamana, Sulaeman Hadi; dan Gata, Windu. "Sentimen Analisis Stay Home menggunakan metode klasifikasi Naive Bayes, Support Vector Machine, dan k-Nearest Neighbor". Jurnal Informatika dan Komputer. Vol.22, Nomor 2, 2020.

[9] Irfani, Faizal Fakhri Irfani; Triyanto, Mohamad; Hartanto, Anggit Dwi; dan Kusnawi. "Analisis Sentimen Review Aplikasi Ruangguru". Jurnal Bisnis Manajemen dan Informatika. Vol. 16, Nomor 3, Februari 2020.

[10] Filemon, Bryan; Mawardi, Viny Christanti; dan Perdana, Novario Jaya. "PENGGUNAAN METODE SUPPORT VECTOR MACHINE UNTUK KLASIFIKASI SENTIMEN E-WALLET". Jurnal Ilmu Komputer dan Sistem Informasi. Vol. 10, Nomor 1, 2022.

Willyanto Wijaya, Seorang mahasiswa program studi Teknik Informatika Universitas Tarumanagara, Jakarta.

Dyah Erny Herwindiati, Memperoleh gelar S.Si dari Institut Teknologi Sepuluh Nopember tahun 1988. Kemudian memperoleh gelar M.Si. dari Institut Pertanian Bogor tahun 1997 dan memperoleh gelat Doktor dari Institut Teknologi Bandung pada tahun 2006. Saat ini aktif sebagai Dekan dan Dosen Teknologi Informasi Teknik Informatika Universitas Tarumanagara, Jakarta.

Novario Jaya Perdana, Memperoleh gelar S.Kom dari Institut Teknologi Sepuluh Nopember tahun 2011. Kemudian memperoleh gelar M.T. dari Universitas Indonesia tahun 2016. Saat ini aktif sebagai Dosen Tetap Perjanjian Teknologi Informasi Universitas Tarumanagara, Jakarta.